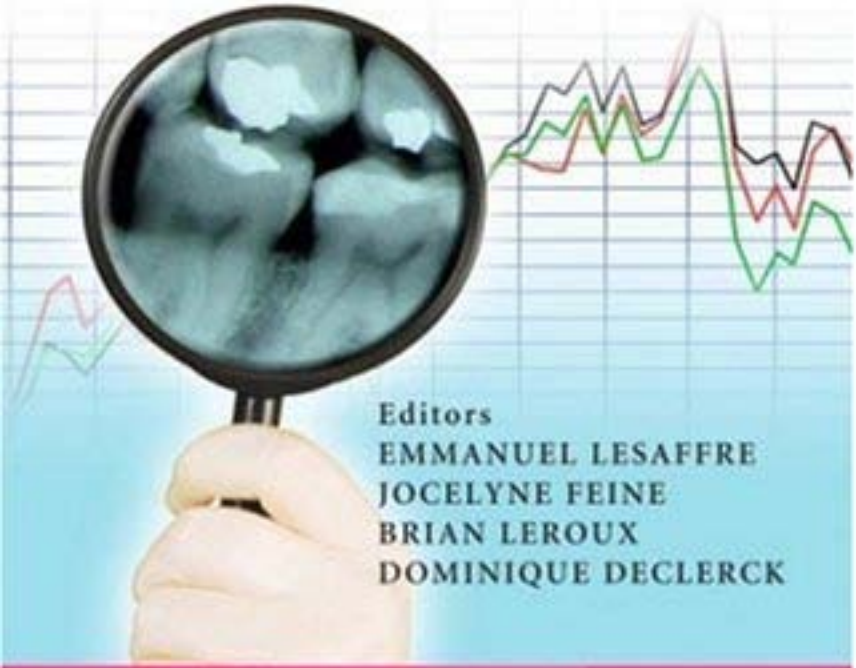
Statistical and Methodological Aspects of Oral Health Research



Editors
EMMANUEL LESAFFRE
JOCELYNE FEINE
BRIAN LEROUX
DOMINIQUE DECLERCK

WILEY

STATISTICS IN PRACTICE

Statistical and Methodological Aspects of Oral Health Research

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

Statistical and Methodological Aspects of Oral Health Research

Edited by

Emmanuel Lesaffre

I-Biostat, Catholic University of Leuven, Belgium and Department of Biostatistics, Erasmus Medical Center, The Netherlands

Jocelyne Feine

Oral Health and Society Research Unit, McGill University, Canada

Brian Leroux

Department of Biostatistics, University of Washington, USA

Dominique Declerck

School for Dentistry,
Catholic University of Leuven, Belgium

A John Wiley and Sons, Ltd., Publication

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Statistical and methodological aspects of oral health research/edited by Emmanuel Lesaffre ... [et al.].
p. ; cm. – (Statistics in practice)
Includes bibliographical references and index.
ISBN 978-0-470-51792-5 (cloth)
1. Dental public health–Research–Statistical methods. 2. Dentistry–Research–Statistical methods.
I. Lesaffre, Emmanuel. II. Series: Statistics in practice.
[DNLM: 1. Dental Research. 2. Health Services Research–methods. 3. Oral Health.
4. Research Design. 5. Statistics as Topic. WU 20.5 S797 2009]
RK52.S67 2009
362.197′60072–dc22
2009004185

A catalogue record for this book is available from the British Library

ISBN: 978-0-470-51792-5

Typeset in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed in the UK by TJ International, Padstow, Cornwall.

Contents

List of Contributors

Finbarr Allen Cork University Dental School & Hospital, Wilton, Cork, Ireland

Todd A. Alonzo Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA

Amy D. Anderson Department of Mathematics, Western Washington University, Bellingham, Washington, USA

Manal A. Awad Department of General and Specialist Dental Practice, College of Dentistry, University of Sharjah, Sharjah, United Arab Emirates and Faculty of Dentistry, McGill University, Montreal, Canada

Christophe Bedos Oral Health and Society Research Unit, Faculty of Dentistry, Montreal, Canada

Melissa D. Begg Columbia University Mailman School of Public Health, New York, USA

Dominique Declerck School for Dentistry, Oral Pathology and Maxillofacial Surgery, Catholic University Leuven, Belgium

Timothy A. DeRouen Departments of Dental Public Health Sciences and Biostatistics, University of Washington, Seattle, USA

Jocelyne S. Feine Oral Health and Society Research Unit, Faculty of Dentistry, McGill University, Montreal, Canada

Stuart A. Gansky Center to Address Disparities in Children's Oral Health, Division of Oral Epidemiology & Dental Public Health, University of California, San Francisco, USA

Thomas A Gerds Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

Peter J. Giannini University of Nebraska Medical Center College of Dentistry, Lincoln, Nebraska, USA

Philippe Hujoel Department of Dental Public Health Sciences, University of Washington, Seattle, USA

ALEJANDRO JARA Departamento de Estadística, Facultad de Ciencias Físicas y Matemáticas Universidad de Concepción, Chile

NIELS KEIDING Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

ARNOŠT KOMÁREK Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

HELMUT KÜCHENHOFF Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany

BRIAN LEROUX Department of Biostatistics, University of Washington, USA

ROOS LEROY Catholic University Leuven, Belgium

EMMANUEL LESAFFRE I-BioStat, Catholic University of Leuven, Leuven, Belgium; and Universiteit Hasselt, Belgium and Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

ALISSA LEVINE Oral Health and Society Research Unit, Faculty of Dentistry, Montreal, Canada

CHRISTINE LOIGNON Oral Health and Society Research Unit, Faculty of Dentistry, Montreal, Canada

DONALD E. MERCANTE, Biostatistics Program, School of Public Health, Souisiana State University Health Sciences Center, New Orleans, LA, USA

NICOLAS NAGELKERKE Department of Community Medicine, Faculty of Medicine and Health Sciences, United Arab Emirates University, United Arab Emirates

IAN NEEDLEMAN International Centre for Evidence-Based Oral Health, Unit of Periodontology, UCL Eastman Dental Institute, London, UK

JOHN M. NEUHAUS Department of Epidemiology & Biostatistics,. University of California, San Francisco, USA

JEFFREY B. PAYNE Department of Surgical Specialties, College of Dentistry, University of Nebraska Medical Center, Lincoln, Nebraska, USA

MARK PEARCE Institute of Health and Society, Newcastle University, UK

CHRISTIAN B. PIPPER Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

PIERRE PLUYE Faculty of Medicine, Department of Family Medicine, McGill University, Montreal, Canada

VIBEKE QVIST Department of Cariology and Endodontics, University of Copenhagen, Copenhagen, Denmark

FAAHIM RASHID McGill University, Montreal, Canada

THOMAS H. SCHEIKE Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

JIMMY STEELE Institute of Health and Society and School of Dental Sciences, Newcastle University, UK

JULIE A. STONER Department of Biostatistics and Epidemiology, College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA

JÖRG R STRUB Department of Prosthodontics, University Hospital, Freiburg, Germany

STEPHANIE D. WOLLIN McGill University, Montreal, Canada

JACKIE VANOBBERGEN Ghent University, Belgium

HELEN WORTHINGTON School of Dentistry, The University of Manchester, Manchester, UK

Preface

The genesis of this book occurred more than ten years ago, when an oral health researcher (DD) invited a biostatistician (EL) to collaborate on a dental longitudinal study. The initiation of the Signal Tandmobiel® longitudinal project instigated the start of a collaboration that continues to the present day. Several PhD dissertations at the Catholic University of Leuven, both in oral health research and (bio)statistics, arose from that early interaction.

Another product of that interaction was a series of international meetings bringing biostatisticians and oral health researchers together. The first meeting, held in 2004 in Leuven, Belgium, focused on follow-up studies, with the third editor as one of the invited speakers. The aim of the meeting was not only to discuss methodological aspects of oral health studies, but even more so to promote collaboration between oral health researchers and biostatisticians. It was also believed that this type of meeting could create a network of oral health researchers and biostatisticians actively collaborating. The idea for the present book was born prior to the second international dental-statistical meeting in Ghent, Belgium, in 2006, where the second editor gave an invited lecture. Several contributors for this book were selected from the network that was established up to then. Our experiences have shown that such a multidisciplinary network is beneficial for oral health research in many ways: by improving on the methodological quality of oral health research, by promoting methodological research in oral health, and by establishing courses on methodology for oral health researchers, among others.

This book is comprised of contributions from oral health researchers and biostatisticians who have longstanding experience in the area. The topics covered in this book were chosen because of their importance to oral health research, but we cannot claim that they are exhaustive. The selection of topics highlights the wide range of methodological aspects that play a role in planning and conducting oral health research. The book is divided into five parts.

Part I contains three chapters which serve as an appetizer for the book. There are frequent complaints about the poor methodological quality of published papers in oral health research.

Chapter 1 reviews some of the problems that are encountered. Of course, these problems are not uniquely related to oral health, but also appear in medical research. However, given the complexity of dental data, methodological errors have particularly high potential for producing misleading conclusions in oral health

research. The first chapter also briefly reviews what can be done to improve upon the quality of oral health research, thereby setting the stage for the subsequent chapters of the book.

Chapter 2 introduces the evidence-pyramid, which is useful for grading the quality of the different sources of evidence. In this system, expert opinion, biological plausibility, bench research, animal studies, and case-series are ranked lowest in quality, whereas controlled systematic experiments in humans (including case-control studies, cohort studies, and randomized controlled trials) are at the top of the pyramid. This chapter highlights the possible dangers of relying too heavily on low-level evidence, particularly on causal thinking, surrogate endpoints and dental devices.

Chapter 3 defines what evidence-based oral health care is and how research data are used in the process. In this respect the chapter treats the methodological aspects of generating research data and explains how to use them effectively in research synthesis. This chapter also reviews aspects of bias and advises the reader on how to properly report research using the CONSORT and related guidelines.

Part II covers a variety of issues in study design and implementation in six chapters.

In **Chapter 4**, a structure is provided to the reader so that she/he can efficiently plan scientifically valid research that incorporates the multi-disciplinary character of oral health research. The reader is guided by means of twenty questions ranging from (Q1): *What is the question of interest, and how does the question translate into a researchable hypothesis?* to (Q20): *Do you have plans for making the study data accessible to others once the planned analyses have been completed and published?* The chapter also deals with the choice of outcome of interest and study design and touches upon some vital statistical questions, ethical issues, writing and using a protocol, finding resources of funding, etc. In summary, it addresses all of the questions that you need to ask yourself when starting up a research study.

Chapter 5 advises the oral health researcher on how to carry out successful clinical studies making use of management principles that also rule commercial projects. The chapter discusses topics on how one can build a multidisciplinary team, seek Institutional Review Board approval, recruit subjects, manage data and set up multicentre studies.

In **Chapter 6**, the reader is introduced to the design, conduct and analysis of randomized clinical trials. The role of regulatory agencies such as the FDA and the EMEA are highlighted. Various study designs are reviewed with emphasis on the special character of dental data.

Design and analysis aspects of epidemiological oral health studies are reviewed in **Chapter 7**, in which the four main types of epidemiological designs are reviewed, along with their strengths and weaknesses: (a) an ecological study (elaborating on the ecological fallacy), (b) a cross-sectional study; (c) a cohort study and (d) a case-control study. Sampling designs commonly used in epidemiological studies, such as multi-stage and stratified sampling, are described.

Chapter 8 introduces the reader to qualitative research. Qualitative research is defined as a ‘process of understanding based on a distinct methodological tradition

of inquiry that explores a social or human problem'. This discipline uses its own research strategies and is complementary to quantitative research, the topic of all of the other chapters in this book. Since qualitative research is based on a different paradigm, its research strategies are also quite different from those of quantitative research. This chapter discusses practical aspects of conducting qualitative research, e.g. how sampling should be done, how observations should be taken, etc.

Chapter 9 focuses on the quality of clinical data, as well as on different aspects of validity for patient self-reported data (e.g. quality-of-life data). Important steps in the collection of data are discussed, such as the steps for data entry, data calibration and how to avoid and treat missing data.

Part III is devoted to standard statistical methodology, useful for the analysis of most medical data and, hence, of interest to oral health research.

Chapter 10 deals with the basic principles of statistics, starting with descriptive statistics and the choice of the most appropriate graphs to summarize collected data. For statistical inference, notions of probability theory are addressed. Classical distributions, such as the binomial, the Poisson and the Gaussian distribution are reviewed. The most important theorem in statistics, the Central Limit Theorem is explained in an intuitive manner, and its implications on statistical inference are illustrated. Concepts related to hypothesis testing, which is the cornerstone of classical statistical inference, such as the significance level, the P-value, the (95%) confidence interval, Type I and Type II error rates and the power of a test are introduced and exemplified. Standard tests comparing two or more than two proportions or means are reviewed, both for unpaired, as well as paired, cases. In addition, statistical concepts that are intensively used in subsequent chapters, such as likelihood principles, are introduced in this chapter. The chapter ends by highlighting the difference between superiority and non-inferiority testing and by stipulating the many misuses of hypothesis testing in the medical and oral health literature.

Regression and correlation approaches used to examine the relationship between an outcome and (more than) one covariate(s) are described in **Chapter 11**. Simple and multiple linear regression are explained, as are procedures that evaluate the appropriateness of the suggested models. Nonlinear-, generalized linear regression models and survival models are discussed in less detail because they are dealt with in later chapters (Part IV).

In **Chapter 12**, methods to assess the accuracy of oral health diagnostic tests are examined. The sensitivity, specificity, and positive and negative predictive values characterize the diagnostic performance of tests with a binary outcome. The Receiver Operating Characteristic (ROC) curve describes the properties of a diagnostic test based on a continuous measure. All of these concepts are explained and exemplified for independent, as well as correlated, data that often occur in oral health research.

While the chapters in Part III cover relatively basic material, the chapters in **Part IV** are more difficult. The reason is that oral health data often show a complex hierarchical structure: multiple surfaces exist on a single tooth, which resides in a mouth that can include up to 32 teeth. If the unit of analysis is surface or tooth,

then classical tests such as those described in Chapter 10 cannot be applied. In this case, the data are clustered, and the statistical tests and/or regression methods need to take into account their correlated nature.

This is done in **Chapter 13** which describes two classical approaches that can be used for the clustered nature of oral health data: the random effects approach and the Generalized Estimating Equations (GEE) approach. Since estimation of the standard errors in complex models could become quite complex, this chapter also illustrates the use of bootstrap methods. In most, if not all, clinical studies researchers are unable to collect all data as planned, resulting in missing data.

Chapter 14 shows the detrimental effect that missing data can have on a statistical analysis, if they are not properly taken into account. Further, the chapter outlines the classical taxonomy of missing data introduced by Rubin and Little in their 1976 seminal paper. Various statistical approaches to deal with the different types of missing data mechanisms are reviewed and exemplified. Special attention is also paid to cases in which the cluster size is itself informative, e.g. when the interest lies in a tooth characteristic, where the number of existing teeth in the mouth is relevant. In many studies the outcome of interest is the time to an event, which is often a failure (of an organ, the occurrence of caries, etc). Such studies need to be analyzed using a technique, called *failure time analysis*.

This is the topic of **Chapter 15**. In this chapter, the concept of censoring and the less well-known concept of truncation are introduced and exemplified. While right-censoring is the most common type of censoring in medical research, left- and interval censoring are regularly encountered in oral health research. Therefore, although the classical methods for right-censored data, such as the Kaplan-Meier technique and the Cox regression are explained, their extensions to left- and interval censored events are also addressed. Other important, but somewhat more advanced, topics treated in this chapter are competing risks models and models that take into account the correlated nature of multiple events that pertain to the same mouth, e.g. when looking at the occurrence of caries on multiple teeth located in the same mouth.

Chapter 16 examines the impact of measurement error and misclassification on estimating parameters of a statistical model. Since special techniques are needed to quantify the effect of inaccurately measuring the covariates and/or the outcome of a regression model, it is not sufficient to simply report a measure of intra- or inter observer variability, such as the kappa statistics, concluding that the measurement process was of sufficient quality. Thus the most relevant techniques are described.

Chapter 17 reviews the basic concepts of statistical genetics. While most of the advances in genetics and genomics research have occurred in medical research, the editors believed that this book would not be complete without dealing with one of the most active areas in health research.

The final chapter of Part III, **Chapter 18**, is devoted to Bayesian methods. The Bayesian approach represents an alternative paradigm to statistical inference, whereby prior knowledge can be incorporated in a formal manner in the statistical analysis. Although applications using the Bayesian approach are presently rather

scarce in oral health research, Bayesian methods have gradually become more important in all empirical research. Thus, the editors believe that Bayesian theory will also become more relevant to oral health researchers.

The book ends with **Part V**, which contains two chapters that extensively describe (a) two intervention studies: Signal Tandmobiel® longitudinal study and the Smile-for-Life study (**Chapter 19**) and (b) a randomized controlled clinical trial (**Chapter 20**). The first study of Chapter 19 has been the basis for many statistical explorations and is used in various chapters of this book for illustration purposes. While in both studies, a simple intervention aims to improve the dietary and brushing behavior of young children, the emphasis in this chapter is not on evaluating the effect of the intervention. Rather, they are provided as detailed examples of well-designed non-randomized oral health interventional studies. The randomized clinical trial treated in Chapter 20 provides detail on all aspects of setting up, conducting and analyzing a clinical trial.

The editors are greatly indebted to four oral researchers who acted as referees for all chapters of the book. More specifically, we wish to thank Roos Leroy, Benjamin Martinez, Kalu Ogbureke and Martha Nunn for their critical but constructive remarks on draft versions of the chapters. The editors especially wish to thank Roos Leroy, who meticulously verified all chapters and provided the authors with ample recommendations for improving their drafts. We also wish to thank Paul Schmitt for commenting on an early version of Chapter 18. Further, thanks also go to Marek Molas and Sten Willemsen, PhD students of the first editor for their help in text processing some of the chapters.

Finally, the first editor wishes to thank his wife Lieve for her patience during the preparation of this book.

The Editors

Part I

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

1

Do we need to improve oral health research?

Dominique Declerck and Emmanuel Lesaffre

1.1 Introduction

The goal of research is to expand our knowledge. Based on the insights obtained, decisions and choices can be made in order to organize our society. The ultimate aim of oral health (OH) research, therefore, must be to accomplish this at the level of oral health and related aspects. Ideally, prevention of disease development, treatment of existing disease and the organization of care delivery should be based on high-level evidence obtained from high-quality research. Statistical and methodological issues determining research quality in the field of OH research are covered in this book.

Research involves different stages, it starts with the planning of the study and it ends with the dissemination of the conclusions. In each of the stages, there is the risk of taking wrong decisions thereby blurring the outcome of the research. In this chapter we elaborate on what might go wrong in the different stages of OH research. We hope that this chapter is an appetizer for the subsequent chapters where most methodological aspects of oral health research will be explained and critically examined. Explicit references to chapters will be made which hopefully will guide the reader throughout the book.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

1.2 Is there a problem?

Several publications can be found reporting on quality issues in different fields of OH research, even in the recent literature. Statistical aspects of design and analysis of randomized controlled clinical trials (RCTs) for the prevention of caries were considered in the paper by Burnside *et al*. (2006). These authors concluded that in recent RCTs of topical fluoride interventions the design of the study was not taken into account. Indeed, while a cluster-randomized design was chosen, the statistical analysis afterwards (see Chapters 6 and 13) most often ignored the clustering of subjects. In addition, essential information was often lacking in the reports not allowing judgement of presence of e.g. possible consent bias. Robinson *et al*. (2006) discussed the quality of reports of RCTs comparing manual and powered toothbrushes. Several important shortcomings were identified: inadequate generation of the randomization sequence, inadequate concealment of the allocation of treatments, inappropriate use of the split-mouth design, etc. A similar finding was reported by Lesaffre *et al*. (2007) on split-mouth trials. Split-mouth studies are popular e.g. in the area of dental materials research. In Lesaffre *et al*. (2007) a review was made of the appropriateness of the statistical approach and on the quality of reporting. Surprisingly, many of the studies did not motivate the use of the split-mouth design and in many papers no full advantage was taken of its design in the statistical analysis. Further, the majority of the studies had a major flaw in their statistical analysis and/or in the reporting of the results. In periodontal research, Tu and co-workers (2006a) evaluated the quality of trials that compared guided tissue regeneration with use of enamel matrix derivatives. Most trials did not meet the majority of their design criteria, so that they placed doubt on the value of and the conclusions from these trials. Eckert and co-workers (2005) presented a literature review on the quality of evidence comparing the clinical performance of dental implant systems. They concluded that the evidence supporting implant therapy was generally derived from case series rather than cohort studies or RCTs. Reports on a direct comparison of different implant systems were not available. In orthodontic research, statistical problems are also common as indicated by Tu *et al*. (2006b).

There are not only problems with the design and the statistics, the problems also exist at the level of taking the measurements. Robinson *et al*. (2006) indicated that plaque data were reported using ten different indices and gingivitis with nine indices. In a recent review of methodological aspects of caries experience screening in epidemiological surveys, Agbaje *et al*. (2009) reported that detailed information was lacking in a considerable number of reports. Moreover, when the use of standardized survey methodology was mentioned (e.g. WHO guidelines), deviations from the original recommendations were often present hampering comparisons between surveys.

1.3 Is oral health research unique?

The problems in quality are not limited to OH research but are commonly encountered in many other fields of medical research. For instance, in a review article by Sjögren and Halling (2002), the quality of RCTs was evaluated in different areas of dental and medical research. Again, the authors concluded that there was a clear need to improve the quality of trial reporting.

Since the methodological issues are not limited to OH research, why should we especially be bothered about the methodology in oral health studies? There are several reasons indeed why we should have a special focus on OH research. The first reason is an obvious but important one, i.e. oral health data are often complex in nature. For instance, in caries research dentists are interested in lesions at tooth surfaces. There are more than 100 tooth surfaces on permanent teeth but, unlike data that are obtained on subject level (stroke, cardiac mortality, weight, etc.), the information that one tooth surface brings to us is not independent of the information from other tooth surfaces in the same mouth. We say that the information from tooth surfaces is 'not independent' or also 'is correlated'. The statistical analysis of simple research questions but involving correlated data is challenging and needs special care and statistical expertise, much more so than most medical research.

Despite the great need, the training in basic methodological and statistical skills at European (and probably also elsewhere) medical faculties is in general rather poor. The situation in dental training is certainly no exception on this trend, on the contrary. The need of methodological expertise is great in OH research.

Another problem is that OH research is somewhat isolated from medical research. Consequently certain trends that pop up in medical research appear much later in OH research. For instance, the CONSORT guidelines (see Chapter 3) were adopted in three medical journals (JAMA, BMJ, Lancet) in 1996 but it was only in 1999 that a dental journal (*British Dental Journal*) adopted the CONSORT policy of reporting. Other examples of a different attitude in medical and oral health research are given in Chapter 2 (cf. surrogate markers).

As medical research, OH research has become increasingly complex with need for multidisciplinarity. The need for technical assistance is obvious and recognized, e.g. when dedicated computer hard- and/or software needs to be developed. Also, when OH research involves new technical innovations, engineers are a necessity. But the need for collaborative efforts with methodologists/statisticians is not sufficiently appreciated. Partly the blame for this attitude is the poor refereeing process on methodology of many dental journals, not forcing the oral researchers to collaborate.

The message is not 'medical research is great and oral health research is poor', but there is no doubt a discrepancy. In the next sections, we review the risks of taking the wrong decisions in performing OH research. We distinguish three stages: (1) planning of the study, (2) conduct of the study and (3) the analysis stage of the

study: statistical analysis of the data, reporting of the results, interpretation of the results and conclusions.

1.4 What to do to improve the quality of oral health research?

1.4.1 Planning stage

An example of poor research is a retrospective study (1) where data is collected from medical records by students in their free available time; (2) without a clear view of what will be investigated; (3) without any idea of how much time and resources are needed to perform the research and (4) what expertise is needed to yield high quality output. All of the necessary steps for performing high quality (OH) research are described in Chapter 4. By means of 20 questions the authors guide the researchers to take the correct steps in planning their research. The authors' emphasis is on setting up OH RCTs, but much of their advice also applies to epidemiological studies.

When considering a RCT or an epidemiological study, the reader is referred to Chapter 6 where the characteristics of a RCT are laid down and to Chapter 7 where the focus is on epidemiological research. The reader is also advised to consult Chapter 2 where reflections are given on what information the different types of research are bringing us.

As explained in Chapter 4, researchers need to reflect on (1) the research question and the associated primary and secondary outcomes, (2) which hypotheses to test, (3) the study design, (4) how to recruit study subjects and how many individuals are needed, (5) ethical aspects, (6) financial aspects, (7) which and how much personnel is needed, etc.

In Chapter 8, the principles of qualitative research are explained. It appears that this kind of research has a lot to offer to the researcher in helping to ask the right kind of questions to the patients, in other words to collect the relevant information for the research questions at hand.

The involvement of statistics/statisticians starts at the planning phase when deciding for the primary endpoint, the choice of statistical analyses, the necessary sample size, etc. Statistical principles are also involved during the conduct of the study, e.g. when an interim analysis of the data is envisaged, but in general statistics are more manifest when planning the study, analyzing the data and interpreting the results. In Chapter 6, a review is given of the different statistical aspects involved in RCTs.

1.4.2 Conduct of the study

In Chapter 5 the authors reflect on building up a multidisciplinary team and managing it to yield a successful outcome. In this respect, the authors elaborate on the basic team management principles. For instance, they reflect on the necessary

qualities of the principal investigator, the co-investigators, the study coordinator, the clinical research monitor, etc. But also aspects of recruitment of patients, how to practically organize randomized allocation and blinding are treated. As in Chapter 4 the authors focus primarily on RCTs, but the chapter is also useful to epidemiological studies.

While data management is becoming more and more computerized, there are still a lot of quality issues that need to be considered in this context. The level of sophistication differs from study to study depending on factors such as the type of study, the financial resources, etc. For example, electronic data entry is becoming a standard in the conduct of RCTs, while in hospital-based studies or small-scale epidemiological surveys data are often recorded on paper and later transferred to the electronic database. Both systems require special attention to guard the quality of the data. Issues that relate to data management problems are also treated in Chapter 5.

A careful planning of the study involves the reflection on which measurements to take. Reasons for not choosing a particular measurement on board are: unclear relationship to the primary questions, (perceived) lack of quality and validity of the measurement process, high likelihood to suffer from missing data, etc. All of these aspects are treated in detail in Chapter 9.

1.4.3 Analysis stage of the study

1.4.3.1 Statistical analysis of the data

Upon collection of the data the exciting part of the research can start, i.e. the statistical analysis of the data leading eventually to the answers to the research questions. However, without a basic understanding of statistics it is impossible to perform any data analysis. Chapter 10 introduces the basic statistical concepts and represents the absolute minimum knowledge that an OH researcher should have for (1) either to do some basic analyses or (2) more importantly, to be prepared for the confrontation with the statistician. This chapter describes a variety of univariate statistical techniques, i.e. methods that analyse each measurement separately. In epidemiological studies, though, it is imperative to correct for imbalances of the risk groups at baseline, leading automatically to regression models. The mother of all regression models is the (multiple) linear regression model where the response is a continuous measurement. Statistical analyses that pertain to this approach are described in Chapter 11 as well as the related correlation approaches. The most popular epidemiological regression models, the logistic regression model and the Cox proportional hazards (survival) model, are also introduced in this chapter.

The statistical implications of the hierarchical structure of oral health data are explained in Chapter 13. Dedicated statistical approaches that deal with such correlated data in an appropriate manner are discussed. These methods were not especially developed for the analysis of oral health data, but they are especially useful for these data. Chapter 15 describes approaches in survival analyses that take into account the correlated nature of the data. The authors also elaborate on

many issues in survival analysis that are especially important in the analysis of survival data.

Coarsened data appear everywhere in medical research, as well as – and perhaps even more so – in OH research. By coarsened data we mean missing data and data that are measured roughly such as the timing of caries development (only known to occur in a time interval; called interval-censored data). The impact and treatment of missing data are treated in Chapter 14. The second type of data are treated in Chapter 15. Less dramatic than missing data, but still problematic, are data that are measured with error. When a (binary, categorical) disease state, e.g. caries experience (Y/N), or a (binary, categorical) risk factor like taking sweets in between meals (Y/N) is recorded with error one speaks of misclassification. Chapter 16 explores what the harm is of measurement error/misclassification and looks for ways to deal with it. Chapter 12 deals with methods that measure the agreement between scoring methods and with methods that measure the predictive ability of imperfect measures vis-à-vis a benchmark examiner or gold standard.

Let it be clear that the application of sophisticated statistical techniques can not correct for a badly planned study or for a poor data collection. This is too often misunderstood by OH researchers but also by medical researchers. For instance, no sophisticated statistical technique can repair (completely) the damage that is done by lacking data (problem of missing data). Oral health data have inherently a complex nature and therefore one cannot always hope for simple analysis methods even to solve simple research questions. Thus the OH researcher is bound to collaborate with statisticians.

We end this section by pointing out that two chapters – Chapter 17 on the analysis of genetic studies and Chapter 18 on the Bayesian methodology – are included to introduce the reader in these popular and rapidly evolving areas.

1.4.3.2 Reporting of the results

Results need to be reported and summarized such that an accurate picture emerges of what data have been collected, what problems were encountered and what statistical methods have been applied. The CONSORT (or similar) guidelines introduced in Chapter 3 help the researcher to write down in detail his/her results.

1.4.3.3 Interpretation of the results and conclusions

The final steps of the research are the interpretation of the statistical analyses and the final conclusions. With a detailed protocol and statistical analysis plan, the freedom in interpreting the results is limited but only for the primary (and secondary) endpoint(s). But such limitations do not hold for all other information that has been collected, or for epidemiological studies with their relatively exploratory character. Moreover, in practice the researchers too often have a biased view upon the results. This is not surprising, since they initiated the research with certain beliefs and hopes. Also, the output of statistical analyses is often only barely understood by clinicians so that opportunities are created to interpret the results in a convenient manner. Examples of such a biased attitude are the phrase ‘there is a trend in

the data' when the P-value is close to 0.05 but still higher, or statements like 'the result would have been definitely significant if the sample size had been greater'. Chapter 2 is in this respect quite useful to obtain an idea of the dangers of making biased conclusions.

Conclusions of the research are typically found in the abstract of the paper and in the Conclusions Section. All of us lack time to read properly our scientific literature. We usually browse through journal issues thereby reading at most the abstract and the conclusions. Also in a further reading of the paper many skip the methodological part of the paper because they lack the technical knowledge to have a good understanding of strengths and weaknesses of the methods applied. It is therefore vital that the conclusions are formulated in a correct manner thereby using the right wording with enough nuances. It is often seen that results are discussed in an appropriate manner in the Results section, but conclusions are too far fetched not supported by the data. Again the CONSORT guidelines can help us in this respect.

1.5 How to assess quality in research?

The assessment of research quality is challenging involving the readers of the paper, but also the referees and the editor of the journal. Especially the referees and the editor have a major responsibility in improving the quality of research. But how can a reader assess the quality of a paper? And what other means do referees and the editor have to judge the quality of reported research?

Indirect assessment of quality A popular way of evaluating the quality of research, especially by the readers, is considering the journal in which the report was published. The Science Citation Index (SCI) was originally introduced in 1961 by the Institute of Scientific Information (ISI). The impact factor of a journal measures the frequency with which the 'average article' in a journal has been cited in a particular year. In this way the overall quality and significance of a given journal's contents are reflected. Journal ranking is increasingly used as a quantitative tool for evaluating journal quality. However, the journal impact factor should only be used as a rough indicator of scientific quality of an individual article (Andersen, 2006). Also, impact factors are greatly influenced by the type of articles (review versus original research), subject speciality and novelty of the research field.

Quality in reporting research Probably the best (we do not claim the perfect, though) tool is the set of guidelines formulated in the CONSORT statement. Several studies reported on the improvement of the quality of reporting in journals where the CONSORT statement was adopted (Moher *et al*., 2001; Kane *et al*., 2007). Mills *et al*. (2005) showed that some recommendations were frequently reported, but reporting of others remained suboptimal. However, it should be kept in mind that possible discrepancies between published reports and actual conduct of randomized clinical trials may still exist (Hill *et al*., 2002). In addition, the quality of

reporting does not guarantee research integrity and cannot protect against scientific misconduct. An example of this is the publication by Sudbo *et al*. (2005) in *The Lancet*, where he stated that the long-term use of non-steroidal anti-inflammatory drugs was associated with a reduced incidence of oral cancer but also with an increased risk of death due to cardiovascular disease. A few weeks after publication of the report some concerns were expressed about this study (Horton, 2006a). It soon became evident that the data had been manipulated: fictitious patients were included in the study. The publication was retracted shortly afterwards and an investigation was started to find out whether other reports by the same author also showed evidence of fraud (Horton, 2006b).

In other research areas, similar initiatives have followed, e.g. for the reporting of meta-analyses of observational studies in epidemiology (MOOSE recommendations – Stroup *et al*., 2000), meta-analyses of randomized trials (the QUORUM statement – Moher *et al*., 1999), diagnostic studies (the STARD initiative – Bossuyt *et al*., 2003) or observational studies in epidemiology (the STROBE statement – von Elm *et al*., 2007).

1.6 Which actions to take

There are a lot of research groups that are doing excellent research based on solid methods. But, there is also a lot of variability in the quality of medical and OH research. We do not expect this to change overnight. If things improve they can only do so gradually whereby first of all the problem should be recognized. Recognition as well as the intention to improve is a must, then comes educational efforts in methodology but also the collaboration with methodologists, not only statisticians. Current research is multidisciplinary, the earlier one realizes this in life the better the research output will be. This implies that efforts should be undertaken to organize joint meetings with oral health researchers and statisticians/methodologists to eventually create networks which will promote collaborative activities and increase the quality of oral health research.

References

Andersen, J., Belmont, J., & Cho, C.T. (2006) Journal impact factor in the era of expanding literature. *J Microbiol Immunol Infect* **39**: 436–43.

Agbaje, O., Lesaffre, E., & Declerck, D. (2009) Assessment of caries experience in epidemiological surveys. Community Dental Health (submitted).

Burnside, G, Pine, C.M., & Williamson, P.R. (2006) Statistical aspects of design and analysis of clinical trials for the prevention of caries. *Caries Res* **40**: 360–5.

Bossuyt, P.M., Reitsma, J.B., & Bruns, D.E., *et al*. (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* **138**: 40–4.

Eckert, S.E., Choi, Y.G., Sanchez, A.R., & Koka, S. (2005) Comparison of dental implant systems: quality of clinical evidence and prediction of 5-year survival. *Int J Oral Maxillofac Implants* **20**(3): 406–15.

Hill, C.L., La Valley, M.P., & Pelson, D.T. (2002) Discrepancy between published report and actual conduct of randomized clinical trials. *J Clin Epidemiol* **55**: 783–6.

Horton, R. (2006a) Expression of concern: non-steroidal anti-inflammatory drugs and the risk of oral cancer. *Lancet* **367**: 196.

Horton, R. (2006b) Retraction – non-steroidal anti-inflammatory drugs and the risk of oral cancer: a nested case-control study. *Lancet* **367**: 382.

Kane, R.L., Wang, J., & Garrard, J. (2007) Reporting of randomized clinical trials improved after adoption of the CONSORT statement. *J Clin Epidemiology* **60**: 241–9.

Lesaffre, E., Garcia-Zattera, M., Redmond, C., Huber, H., & Needleman, I. (2007) A review of the reported methodological quality of split-mouth studies. *Journal of Clinical Periodontology* **34**: 756–61.

Mills, E.J., Wu, P., Gagnier, J., Devereaux, P.J. (2005) The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. *Contemp Clin Trials* **26**(4): 480–7.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D.F. (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. Quality of Reporting of Meta-analyses. *Lancet* **354**: 1896–1900.

Moher, D., Jones, A., & Lepage, L. (2001) Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* **285**: 1992–5.

Robinson, P.G., Damien Walmsley, A., Heanue, M., *et al*. (2006) Quality of trials in a systematic review of powered toothbrushes: suggestions for future clinical trials. *J Periodontol* **77**(12): 1944–53.

Sjögren, P. & Halling, A. (2002) Quality of reporting randomized clinical trials in dental and medical research. *Br Dent J* **192**(2): 100–3.

Stroup, D.F., Berlin, J.A., Morton, S.C., *et al*. (2000) Meta-analysis of observational studies in epidemiology. A proposal for reporting. *JAMA* **283**(15): 2008–12.

Sudbø, J., Lee, J.J., Lippman, S.M., *et al*. (2005) Non-steroidal anti-inflammatory drugs and the risk of oral cancer: a nested case-control study. *Lancet* **366**: 1359–66.

Tu, Y.K., Maddick, I., Kellett, M., Clerehugh V, & Gilthorpe MS. (2006a) Evaluating the quality of active-control trials in periodontal research. *J Clin Periodontol* **33**(2): 151–6.

Tu, Y.K., Neslon-Moon, Z.L., & Gilthorpe, M.S. (2006b) Misuses of correlation and regression analyses in orthodontic research: the problem of mathematical coupling. *Am J Orthod Dentofacial Orthop* **130**(1): 62–8.

von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C., & Vandenbroucke, J.P. for the STROBE initiative (2007) The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* **370**: 1453–7.

2

Grading evidence with a focus on etiology, surrogates, and clinical devices

Philippe Hujoel

2.1 Introduction

'I think two and two is four, not three or five. I'm an evidence-based guy. We are trapped in a reality-based world.' Bill Clinton, Financial Times, Monday, 24 September, 2007

We all live in an evidence-based world where $2 + 2$ should equal 4, not 5 or 3. Unfortunately, judging the soundness of evidence in medicine or dentistry is not as straightforward as a simple addition. On many clinical questions, available evidence is arguably incomplete, conflicting, or nebulous, making it difficult to arrive at the yes/no decision required in daily clinical practice. In choosing a restorative material, how should one judge the common claim that gold is the 'standard' restoration because gold has high corrosion resistance or malleability, versus the clinical trial evidence that the failure rate of gold cast restoration is 10 % (Hayashi & Yeung, 2003)? In 1992, a landmark article proposed an evidence-based approach to medicine which addressed such challenges (Guyatt, 1992). One important aspect of this evidence-based approach is the recognition that in medicine, and in dentistry, there are formal rules in evaluating the reliability of evidence.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

While it would be attractive to surmise that these formal rules for grading evidence were based on astute thinking or logic, they are unfortunately often the result of prior disasters and the laws, regulations, and guidelines that followed the disasters. A benign sulfa drug dissolved in toxic diethylene glycol led to over a hundred deaths – mostly children – and precipitated the Federal *Food, Drug, and Cosmetic Act of 1938*. Drug manufacturers were for the first time mandated to provide evidence of drug safety prior to marketing. The thalidomide crisis led to the 1962 *Kefauver-Harris Amendments*, which have been credited with the increased utilization of randomized controlled trials to provide evidence on effectiveness. Deaths associated with a device – the Dalkon Shield® (Sivin 1993) – precipitated the *Medical Device Amendment* of 1976 which requires evidence that high-risk devices (e.g. angioplasty catheters) are safe and effective. Possibly, the controversies surrounding approved drugs such as encainade, rofecoxib, or rosiglitazone may lead to further changes in the methods and principles of evaluating clinical evidence Psaty *et al*., 1999). Such cycles of disasters and responses led to the recognition of a hierarchy of evidence upon which to judge drugs and devices. Evidence-based medicine came along and formalized rules in assessing the clinical literature not only on drugs and devices, but also clinical procedures. The advent of evidence pyramids or evidence-levels has helped in grading and translating evidence into a clinical recommendation.

In the evidence-pyramid, the lowest levels of evidence are expert opinion, biological plausibility, bench research, animal studies, and case-series (Figure 2.1). Such findings are relegated to the lowest tier of evidence not because such research or thinking is methodologically unsound or because the evidence derived is unreliable. Quite the opposite is true. Animal research has been reported to form the basis for two-thirds of the Nobel prizes in Physiology or Medicine (AMP News Service 2008). Bench research has led to discoveries such as the structure of DNA, the discovery of antibiotics, or the synthesis of insulin. Case-series, with implicit historical controls, led to the identification of bisphosphonates as a cause for osteonecrosis of the jaw or to a Nobel Prize in Physiology or Medicine for discovering bone marrow transplantation as a treatment for leukemia. And finally, biological plausibility is at the basis of treatments such as root canals for the treatment of acute pulpitis. The adjective ‘low-level’ does not refer to the intrinsic (or inherent) quality or value of the evidence, but rather as to how the evidence is valued when it is used as a basis for making clinical decisions for humans.

The term ‘low-level’ is used because most commonly leaps of faith are required to assume that knowledge on biological mechanisms, or results obtained from animal or bench experiments translate into clinical decisions that lead to a tangible patient benefit. Medical history has shown that such leaps of faith have an unacceptably high chance of leading to harmful clinical decisions (think ‘disasters’). For instance, a vaginal cream effective at preventing HIV transmission in macaques unexpectedly increased HIV infection rates in humans (Van Damme *et al*., 2002). Or, omeprazole, one of the most widely prescribed gastric proton pump inhibitors internationally, increases the incidence of gastric carcinomas in rats (Ekman *et al*., 1985), but without such evidence in humans. Benefits or harms in animals may

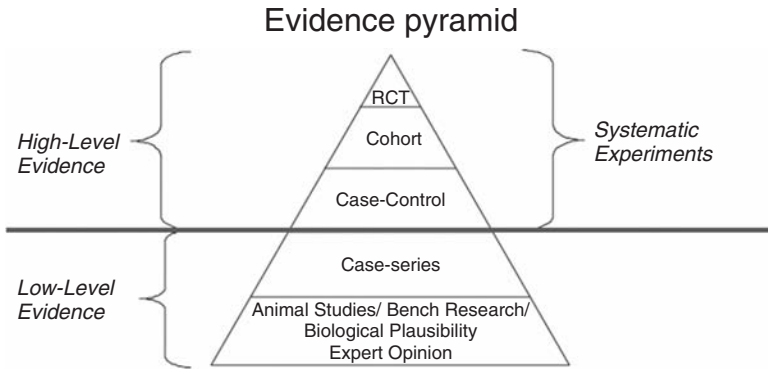


Figure 2.1 The evidence-pyramid ranks the reliability of evidence from low (bottom tier of pyramid) to high (RCT = randomized controlled trial). The three highest level of evidence consist of systematic experiments and became routine in the second half of the 20th century. With an inverted evidence pyramid, the reliability of evidence is ranked in a reverse order and biological plausibility is considered most reliable and RCTs are 'used' to prove biological plausibility.

not translate to humans. One physician-epidemiologist – Feinstein – has suggested that one of the historical hallmarks of the 20th century is the demise of low-level evidence and the advent of controlled human observations.

High-level evidence consists of controlled systematic experiments in humans: primarily case-control studies, cohort studies, and randomized controlled trials. Instead of using deductive reasoning to connect a cause to its effect in humans, observations are made on a sample of subjects and are generalized using inductive inference. In a case-control study, individuals with and without a disease or a condition are compared with respect to the prevalence of a suspected etiological factor. For instance, individuals with and without brain cancers can be compared with respect to the prevalence of past medical and dental diagnostic X-ray exposures. In a cohort study, exposed and non-exposed individuals are followed longitudinally and the incidence of the outcome of interest is monitored. For instance, individuals exposed and not exposed to water fluoridation can be followed longitudinally and monitored for the incidence of osteosarcomas. Both case-control studies and cohort studies are commonly used to elucidate causes of disease.

Health recommendations which are based on such epidemiological study designs (i.e. case-control and cohort studies) can be disastrous. Both hormone replacement therapy and vitamin A supplementation provide two textbook examples of treatments that were touted as saving lives, based on epidemiological studies, yet randomized controlled trials proved them to be harmful. A fundamental problem with epidemiological studies is that both selecting individuals into studies in an unbiased fashion and controlling for the lifestyles and environment of these individuals is to a certain extent impossible and will always lead to biases. Unequivocal control for such biases can only be achieved by a study design referred

to as a randomized controlled trial, the top of the evidence pyramid. In a randomized controlled trial, individuals are randomly assigned to exposures (most commonly treatments) and the incidence of the outcome of interest is monitored. For instance, individuals can be randomly assigned to either xylitol gum or sorbitol gum and the incidence of caries can be monitored. Many of the inherent and often uncontrollable biases present in case-control studies and cohort studies can be eliminated by the powerful, but delicate, mechanism of randomly assigning individuals to treatments. Certain evidence-based organizations such as the Cochrane Collaboration restrict the evidence to randomized studies when reporting on treatment effectiveness.

These three different types of systematic experiments (case-control studies, cohort studies, and randomized trials) are powerful tools to evaluate whether the low-level evidence on cause and effect indeed translates into tangible outcomes in humans. While there has been a rapid rise of high-level evidence studies in dental research, the published research is sometimes in violation of the basic principles of statistics and epidemiology that form the foundation for reliable results of high-level evidence. Examples include reports using two-sample t-tests for split-mouth trials (Lesaffre *et al*., 2007), basing almost all clinical research on unvalidated surrogate endpoints (Hujoel, 2004), or failing to identify how controls are selected in case-control studies (Lopez *et al*., 2007).

Why is it that some of our dental clinical decisions remain based on poor-quality evidence? The common lack of high-level evidence for important clinical decisions may have multiple causes including lack of resources, conflicts of interests, or a belief that low-level evidence trumps high-level evidence. This latter concern will be referred to as the problem of the inverted evidence pyramid; the belief that formal rules for grading evidence that exists in evidence-based medicine do not apply to dental decision making; that high-level evidence is unnecessary because the clinical practices of diagnosis, prognosis, and treatment are so obviously effective that requesting clinical research – high-level evidence – is unethical or a waste of resources. For those who adhere to an inverted evidence pyramid, high-level evidence is considered not necessary, inferior to biological plausibility and the conduct of clinical research – if needed at all- is a perfunctory task to prove obvious biological truths. This chapter will focus on the pernicious role of low-level evidence on causal thinking, surrogate endpoints, and devices.

2.2 The missing epidemiological baton on causality

The Framingham Heart Study, which has become an icon for modern causal thinking (Levy & Brink, 2005), is an example of the superiority of high-level evidence on causality. Prior to this study, low-level evidence suggested that women were immune to heart disease and that high blood pressure was a normal physiological response to aging. The Framingham heart study refuted these hypothesized causes, which were based on low-level evidence, and instead used a cohort study design to identify cigarette smoking and blood pressure as causes of cardiovascular disease.

The Framingham Heart Study exemplified the power of high-level evidence in helping to understand the causality of complex chronic diseases. The role of epidemiology in the big scheme of medical discovery solidified. 'Medical science (in cardiovascular disease studies) continually passes the baton of discovery from (epidemiological) observation to laboratory studies to human clinical trials' (Levy & Brink 2005). In cardiovascular disease research, epidemiology identified the risk factors such as high blood pressure, basic scientists explored drugs that lowered blood pressure, and randomized controlled trials closed the causal loop by showing improved survival with blood pressure medication.

The epidemiological baton of discovery led to the use of fluorides in dentistry, which is considered one of the ten greatest health milestones of the 20th century by the Centers for Disease Control. Between 1901 and 1929, the water supply in a Colorado town was linked to mottled teeth, mottled teeth were linked to caries resistance, and fluoride was identified as the responsible ingredient in the water supply. The baton of discovery passed from epidemiological observations, to the basic science of fluoride biology, to randomized trials showing that fluoride reduces caries rates (Centers for Disease Control and Prevention, 2000). The causal associations between fluoride and caries, and between blood pressure and heart disease, are rooted in the high-level evidence of both observational epidemiology and randomized controlled trials. This high-level evidence on causality contrasts with causal associations that remain rooted in biological plausibility or animal experiments. Unlike blood pressure and heart disease, fluoride and caries, there are several examples in dental research where causal thinking remains strongly influenced by low-level evidence such as pathophysiological reasoning on assumed disease mechanisms, rather than high-level evidence such as can be derived from cohort studies. For instance, one example illustrative of low-level evidence is that infection leads to destructive periodontal disease.

No pivotal cohort study identified infection as the cause of destructive periodontal disease. One citation classic in support of the periodontal infection theory is a case-series of nine dental students, a dental professor, and two dental technicians who brought their gingiva to an unnatural inflammation-free condition referred to as *'Aarhus superhealthy gingiva'* (Theilade, 1987), and then subsequently stopped brushing their teeth (Löe *et al*., 1965). The *experimental* gingivitis that developed in ten to twenty days formed 'didactic proof of the essential role of dental plaque bacteria in periodontal diseases' (Theilade 1987). Findings on *acute experimental* gingivitis were extrapolated to *chronic clinical* gingivitis, and – a huge leap – to periodontal diseases.

In another citation classic in support of periodontal infection, *Bacteroides gingivalis* was injected around ligated teeth in eight female monkeys causing a 'burst' of bone loss (Holt *et al*., 1988). Such low-level evidence led to the conclusion that 'this microorganism (is) of great importance to the control of periodontitis'. This study led to extensive investments into vaccines, microbial diagnostics, and antibiotics. This occurred even though contemporary reviews indicated the absence of prospective etiological studies (high-level evidence) on the role of *Bacteroides gingivalis* in humans (Haffajee & Socransky, 1994). In 1996, the evidence for

Bacteroides gingivalis was described as 'disappointing' at a World Workshop conference (Offenbacher, 1996), and more recent cohort studies failed to identify this organism's role in destructive periodontal disease (Dahlen *et al*., 1995, Tanner *et al*., (1998), Timmerman *et al*., (2000), Machtei *et al*., (1999)). The baton of discovery on *Bacteroides gingivalis* never started with epidemiology and the price paid for this lack of epidemiological evidence may have been high[1] in terms of misdirected basic science research, overuse of antibiotics in the clinic, and failure to recognize the true drivers of destructive periodontal disease.

The need for high-level evidence on causality cannot be made without some caveats. First, dental research is not alone in lacking a rigorous high-level basis of causality for some key questions. Investigators in other disciplines such as psychiatry have similarly voiced concerns about the need for higher level evidence: the need 'for a big push from an effort like the Framingham Heart Study' (Helmuth, 2003). Second, epidemiology can provide wrong leads to basic science and intervention studies such as happened with the hormone replacement therapy (Nelson *et al*., 2002) or beta-carotene (Heinonen *et al*., 1994). Finally, most insights in causality, such as fluoride and caries, are often initially based on low-level evidence: 'the experience of seeing, thinking about, and treating individual patients' (Rees 2002). Such low-level evidence is of critical importance to drive the important hypotheses in clinical research. The main point elaborated here is that these hypotheses should then be submitted to the rigorous systematic experimentation that leads to high-level evidence.

2.3 The biological justification of surrogate endpoints

Many dental treatments on the world market are approved based on the argument that subtle changes in dental disease markers translate into tangible patient benefit. A fraction of a millimeter improvement in tissue position, a fraction of milligram change in enamel or bone mineral density is postulated, based on biological plausibility, to translate into reduced tooth loss or increased oral-health-related quality of life. The terminology, issues, and criteria to obtain evidence better than biological plausibility came to the forefront in 1989 with the publication of the *Cardiac Arrhythmia Suppression Trial* (Anonymous, 1989). Two drugs which were effective at suppressing symptomatic ventricular arrhythmia, and therefore believed to

[1]In contrast to the ever-changing picture on the role of infection, the emerging epidemiological baton on the causality of destructive periodontal disease points consistently to smoking and possibly unhealthy eating habits. Current high-level evidence suggests that 50% to 80% of all destructive periodontal disease cases are attributable to smoking (Haber *et al*., (1993), Thomson *et al*., (2007), Tomar & Asma (2000)). Yet, the infection theory, low-level evidence, suppressed the epidemiological evidence on smoking (Bergstrom, 2004). In 1990, a review in the *New England Journal of Medicine* still did not identify smoking as a risk factor (Williams, 1990). The inverted evidence pyramid may have obscured the primary drivers of the destructive periodontal disease epidemic, funneled resources in a 'disappointing' direction, and focused patient-care on bacteria such as *Bacteroides gingivalis* which turned out to be of dubious importance.

improve survival, turned out to substantially increase mortality risk. An objective sign derived from an electrocardiogram (i.e., suppression of ventricular arrhythmias) provided misleading information when it came to assessing whether the treatment improved survival: an unequivocal sign of tangible patient benefit. A good therapeutic result on an electrocardiogram translated into a dead patient.

Subsequent to this 1989 landmark study, the term surrogate became common parlance and was defined by the Food and Drug Administration as: 'a laboratory sign or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives. Changes induced by therapy on a surrogate endpoint are expected to reflect changes in a clinically meaningful endpoint' (Temple, 1995). More succinctly, a true endpoint reflects unequivocal evidence of tangible patient benefit (Fleming, 2005), whereas a surrogate endpoint reflects an intangible patient benefit.

From the beginning there was a contrast between dentistry and medicine in the evidence available to justify the use of surrogate endpoints in clinical studies. In medicine, cohort studies had related surrogates such as blood pressure, and ventricular arrhythmias to subsequent mortality: the predictiveness of those surrogates selected as endpoints in randomized controlled trials was rooted in high-level evidence. In dentistry, because of the missing epidemiological baton, there was minimal or no evidence from cohort studies to select surrogates. Surrogates are reported to be strong predictors of a true endpoint based on biological plausibility arguments such as 'large losses ... could lead to tooth loss', not because epidemiological studies suggested small changes predicted 'large losses', let alone a tangible outcome.

Research on surrogates in AIDS and cancer provided evidence-based guidelines to improve the likelihood that a surrogate endpoint is valid; i.e. that it may provide the correct conclusions in the assessment of treatment efficacy (Burzykowski *et al*., 2005). Several drug disasters had indicated that predictiveness is, in and of itself, weak evidence to justify a surrogate. 'A correlate does not a surrogate make' (Fleming & DeMets, 1996). One important guideline, adhered to by regulatory organizations, is that the validity of a surrogate marker depends on the closeness to the true endpoint. If the natural disease history is represented by a river, where the river source represents the first subtle subclinical disease marker and the river mouth represents the clinical endpoint, then downstream surrogate endpoints are preferable to upstream surrogate endpoints (Figure 2.2) (Chakravarty, 2005). The first sign of demineralization, the first presence of cancer cells, and the first signs of atherosclerosis are less likely to be valid than more downstream surrogate endpoints such as frank cavitations, cancer metastasis, and 90 % coronary stenosis. By the same guideline, 10 mm of attachment loss is a more favored surrogate than a 1 mm loss.

While drug regulators suggest looking downstream, obtaining high-level evidence on surrogate validity, some dental researchers argue looking upstream, sticking with low-level biological plausibility for claiming surrogate validity, or side-stepping the scientific issues on surrogate endpoints by providing an unique dental definition of surrogate. For instance, panels of dental experts concluded that

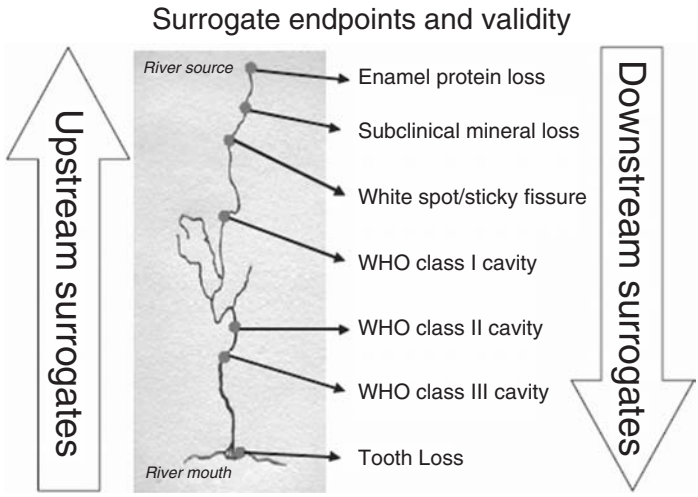


Figure 2.2 River reflecting the natural history of the caries process starting with the first molecular marker of caries (enamel protein loss) and ending with tooth loss. Downstream surrogate endpoints are more likely to identify treatments that provide tangible patient benefits than upstream surrogate endpoints. While the scientific evidence to look downstream is sufficiently convincing for regulatory organizations, dental research often appears to look further upstream, rather than downstream.

'early caries lesions/white spots are NOT a surrogate' (Pitts & Stamm (2004), Pitts (2005)) as if a subclinical loss of some micrograms in tooth mineral density – only detectable by machines that go 'ping' (Doctor Spenser in the movie *The Meaning of Life* (1983)) – is a tangible patient outcome. Defining statistical terminology in such a unique way precludes assessing biological beliefs with high-level evidence. Ironically, much of dentistry has been a surrogate-driven research field where the significance of small surrogate changes often failed to resolve controversies. In the periodontal arena, a NIDCR director stated: 'In spite of attempts in the last 25 years to assess the relative benefits of surgical and non-surgical treatment of periodontitis, there is still no clear and unambiguous answer to this question' (Löe, 1997). Clinical trials with even more upstream surrogates, as some are currently arguing for, are even less likely to resolve controversies.

2.4 Parachutes, radiant smiles, biological plausibility and devices and procedures

The reliance on low-level evidence in the area of devices and procedures has the potential to have serious adverse public health consequences. In the 1980s, the low-level evidence of biological plausibility and expert opinion led to the claim that

'a new era in (Proplast-Teflon) TMJ reconstruction has begun, resulting in increased benefits to the patients whom we all serve' (Moriconi *et al*., 1986). The biological plausibility failed to deliver on these promises of tangible patient benefits. Many of the 26,000 patients with TMJ implants reported disastrous outcomes which in turn led to lawsuits, intense media coverage, and the destruction of government-recalled TMJ implants by way of bulldozer (Mendenhall, 1995). Interestingly, the biological plausibility on TMJ implants was not backed up by animal experiments prior to marketing – in and of itself one of the lowest forms of evidence available.

Hundreds of millions of individuals are exposed to dental materials and products, with often only the lowest level evidence as justification for safety. During one period in the 20th century, uranium was added to dental porcelain, a commonly used material for crowns, under the belief it would provide the recipient with a radiant smile (Thompson, 1976). While the ionizing radiation of the uranium exceeded in some instances the legal limits, it was opined not to be a problem (Nally, 1969). The past use of asbestos in surgical dressing (Otterson & Arra, 1974), and the present use of formaldehyde, another human carcinogen (International Agency for Research on Cancer, 2004), in children (Myers *et al*., 1978) makes it surprising that the safety of dental devices remains, up to this day, largely based on low-level evidence.

Biological plausibility has been wildly successful for some common dental procedures and devices providing a powerful argument in favor of the inverted evidence pyramid. Tooth extraction resolves the pain of acute pulpitis with high effectiveness, and placebo-controlled trials are justifiably uncalled for. Restorative dentistry almost always succeeds at restoring function and esthetics to badly broken down dentitions. These victories in the power of common sense, bench research, and biological plausibility in justifying procedures and devices may have in part provided a way for more questionable interventions to sneak in without the support of high-level evidence.

Indeed, the common sense logic behind some dental procedures appears so convincing that requesting higher level evidence may seem just as preposterous as demanding controlled evidence that parachutes save lives (Smith and Pell, 2003). While the parachute argument should not be disregarded, one must not forget that the biological plausibility of bloodletting was so convincing that it managed to survive for centuries. In dentistry, the benefit of trephination, a common sense procedure, was similarly considered unequivocal for centuries. One textbook stated:

> Trephination in the Western Hemisphere is at least as old as Indian civilization. This surgical form is used to secure drainage and alleviate pain when exudates in the cancellous bone is dammed up behind the cortical plate. The tremendous pressure leads to excruciating pain of acute apical periodontitis ... Scoring the bone with a heavy punch ... speeds relief and healing (Ingle & Beveridge, 1976).

After two double-blind randomized controlled trials on the trephination procedure (Houck *et al*. (2000), Nist *et al*. (2001)), both the safety and the effectiveness were described somewhat differently in a subsequent edition:

> This is a limited-use procedure and is fraught with peril and potential negative complications. . . . Literature pertaining to this procedure is very limited and consists primarily of case reports, opinions, and clinical experiences (Ingle and Bakland, 2002).

Parachute arguments have to be used carefully; low-level evidence, which justified trephination for centuries, was brought into question with the first high-level evidence studies.

One area in which the parachute argument should be regarded with particular suspicion is procedures and devices in which the benefits are uncertain and the harms appear plausible, or where healthy individuals (in this instance, individuals without dental diseases) are advised to modify lifestyles or to use preventive medications such as drinking fluoridated water. Under such circumstances, the highest level of evidence in terms of safety and effectiveness needs to be present. One widely prevalent procedure in this respect is diagnostic radiation where low-level evidence remains the primary justification for many dental X-ray screening programs. Five scientific organizations concluded that low-level ionizing radiation exposures increases cancer risk (National Research Council (2006), Valentin (2005), National Council on Radiation Protection and Measurements (2001), United Nations (2000), Cox *et al.* (1995)). The conclusions of these organizations imply that exposing millions to dental X-rays may be worthwhile if an acceptable harm-benefit ratio is present. For instance, use of diagnostic radiation to screen for colon cancer may be worthwhile if the number of lives saved by intervention exceeds the number of cancers caused by the ionizing radiation (Brenner & Georgsson, 2005). Currently, some of the most expensive government sponsored medical trials are investigating the value of ionizing radiation in screening for lung cancer (Kaiser, 2006). No such trials exist to provide high-level evidence on the benefits of dental X-rays. Reliable answers, high-level evidence, to questions on the benefit of X-rays are not available in part because some clinicians believe that the benefits are so unequivocal, that asking for high-level evidence is unjustifiable. Such arguments should be treated skeptically whenever the evidence on benefit is low-level and evidence for harm is high-level. Devices and procedures should under those circumstances be submitted to the same systematic experiments as drugs to assess safety and efficacy.

2.5 Summary

An inverted evidence pyramid has the potential to lead to arcane causal thinking, bring ineffective drugs, procedures, and devices to the market, and harm dental patients. While the science of systematic experiments is elegant and powerful, it is fragile machinery which can easily be manipulated to ‘prove’ biological plausibility. Surrogate endpoints, subgroup analyses, and contrived comparisons are examples of statistical trickery which can be used to subvert the delicate methods available to rigorously assess biological beliefs. An inverted evidence-pyramid may even make such trickery seem appropriate. However, science is about formulating a biologically plausible hypothesis and then testing the

hypothesis with the best available experimental designs that lead to high-level evidence. The opposite, torturing the world of systematic experiments to make the conclusions fit the low-level evidence of biological plausibility, is a product of the inverted evidence-pyramid.

References

AMP News Service (2008) *Nobel Award in Medicine Holds Strong Message for Animal Activists*. URL: http://www.amprogress.org/site/apps/nl/content2.asp?c=jrLUK0PDLoF&b=1311499&ct=4512083.

Anonymous (1989) Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *N Engl J Med* **321**(6), 406–12.

Bergstrom J (2004) Tobacco smoking and chronic destructive periodontal disease. *Odontology* **92**(1), 1–8.

Brenner D & Georgsson M (2005) Mass screening with CT colonography: should the radiation exposure be of concern?. *Gastroenterology* **129**(1), 328–37.

Burzykowski T, Molenberghs G & Buyse M 2005 *The Evaluation of Surrogate Endpoints*. Statistics for Biology and Health. New York: Springer.

Centers for Disease Control and Prevention (2000) Achievements in public health, 1900–1999: fluoridation of drinking water to prevent dental caries. *JAMA* **283**(10), 1283–6.

Chakravarty A (2005) *Regulatory Aspects in Using Surrogate Markers in Clinical Trials*. New York: Springer.

Cox R, Muirhead C & Stather, J. W. *et al*. (1995) *Risk of Radiation-Induced Cancer at Low Doses and Low Dose Rates for Radiation Protection Purposes*. Documents of the NRPB Vol. 6, No. 1.

Dahlen G, Luan W & Baelum, V. *et al*. (1995) Periodontopathogens in elderly Chinese with different periodontal disease experience. *J Clin Periodontol* **22**(3), 188–200.

Ekman L, Hansson E & Havu, N. *et al*. (1985) Toxicological studies on omeprazole. *Scand J Gastroenterol Suppl* **108**, 53–69.

Fleming T (2005) Surrogate endpoints and FDA's accelerated approval process. *Health Aff (Millwood)* **24**(1), 67–78.

Fleming T & DeMets D (1996) Surrogate end points in clinical trials: are we being misled?. *Ann Intern Med* **125**(7), 605–13.

Guyatt G (1992) Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA* **268**(17), 2420–5.

Haber J, Wattles J & Crowley, M. *et al*. (1993) Evidence for cigarette smoking as a major risk factor for periodontitis. *J Periodontol* **64**(1), 16–23.

Haffajee A & Socransky S (1994) Microbial etiological agents of destructive periodontal diseases. *Periodontol 2000* **5**, 78–111.

Hayashi M & Yeung C (2003) *Ceramic inlays for restoring posterior teeth*. Cochrane Database Syst Rev 2003(1) CD003450.3.

Heinonen O, Huttunen J and Albanes, D. *et al*. (1994) The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. the Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. *N Engl J Med* **330**(15), 1029–35.

Helmuth L (2003) In sickness or in health?. *Science* **302**(5646), 808–10.

Holt S, Ebersole J & Felton, J. *et al*. (1988) Implantation of bacteroides gingivalis in non-human primates initiates progression of periodontitis. *Science* **239**(4835), 55–7.

Houck V, Reader A & Beck, M. *et al*. (2000) Effect of trephination on postoperative pain and swelling in symptomatic necrotic teeth. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* **90**(4), 507–13.

Hujoel P (2004) Endpoints in periodontal trials: the need for an evidence-based research approach. *Periodontol 2000* **36**, 196–204.

Ingle J & Bakland L (2002) *Endodontics (5th edition)*. Hamilton, Ont. London: B C Decker.

Ingle J & Beveridge E (1976) *Endodontics (2nd edition)*. Philadelphia: Lea & Febiger.

International Agency for Research on Cancer (2004) *IARC Classifies Formaldehyde as Carcinogenic to Humans*. URL: http://www.iarc.fr/ENG/Press_Releases/archives/pr153a.html (Accessed 19 August 2007).

Kaiser J (2006) Cancer research. After regime change at the National Cancer Institute. *Science* **312**(5772), 357–9.

Lesaffre E, Garcia-Zattera M & Redmond, C. *et al*. (2007) Reported methodological quality of split-mouth studies. *J Clin Periodontol* **34**(9), 756–61.

Levy D & Brink S (2005) *A Change of Heart: How the Framingham Heart Study Helped Unravel the Mysteries of Cardiovascular Disease (1st edition)*. New York: Knopf.

Löe H (1997) Closing remarks: trials and tribulations. *Annals of Periodontology* **2**(1), 359–63.

Löe H, Theilade E & Jensen S (1965) Experimental gingivitis in man. *J Periodontol* **36**, 177–87.

Lopez R, Scheutz F, Errboe M & Baelum V 2007 Selection bias in case-control studies on periodontitis. A systematic review. *European Journal of Oral Sciences* **115**, 339–43.

Machtei E, Hausmann E & Dunford, R. *et al*. (1999) Longitudinal study of predictive factors for periodontal disease and tooth loss. *J Clin Periodontol* **26**(6), 374–380.

Mendenhall S (1995) Tmj implants: lessons for all of us. *Orthopedic Network News* **6**(2), 1–6.

Moriconi E, Popowich L & Guernsey L 1986 Alloplastic reconstruction of the temporomandibular joint. *Dent Clin North Am* **30**(2), 307–25.

Myers D, Shoaf H & Dirksen, T. R. *et al*. (1978) Distribution of 14C-formaldehyde after pulpotomy with formocresol. *J Am Dent Assoc* **96**(5), 805–13.

Nally J (1969) Uranium content of some dental porcelains and beta activity. *Helv Odont Acta* **13**, 32–5.

National Council on Radiation Protection and Measurements (2001) *Evaluation of the Linear-Nonthreshold Dose-Response Model for Ionizing Radiation*. Bethesda, Md.: National Council on Radiation Protection and Measurements.

National Research Council (2006) *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2*. The National Academies Press.

Nelson H, Humphrey L & Nygren, P. *et al*. (2002) Postmenopausal hormone replacement therapy: scientific review. *JAMA* **288**(7), 872–81.

Nist E, Reader A & Beck M (2001) Effect of trephination on postoperative pain and swelling in symptomatic necrotic teeth. *J Endod* **27**(6), 415–20.

Offenbacher S (1996) Periodontal diseases: pathogenesis. *Ann Periodontol* **1**(1), 821–78.

Otterson E & Arra M (1974) Potential hazards of asbestos in periodontal packs. *J Wis Dent Assoc* **50**(11), 435–438.

Pitts N (2005) *Indiana Conference ICDAS Workshop Group Report. Caries Proceedings of the 7th Indiana Conference, Indianapolis, Indiana, by George K. Stookey and Indiana University School of Dentistry*. ISBN 0965514951/9780965514958/0-9655149-5-1, Publisher Indiana University School of Dentistry.

Pitts N & Stamm J (2004) International consensus workshop on caries clinical trials (icw-cct)–final consensus statements: agreeing where the evidence leads. *J Dent Res* **83**(Spec Iss C), C125–8.

Psaty B, Weiss N & Furberg, C. D. *et al*. (1999) Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. *JAMA* **282**(8), 786–90.

Rees J (2002) Complex disease and the new clinical sciences. *Science* **296**(5568), 698–700.

Sivin I (1993) Another look at the Dalkon Shield: meta-analysis underscores its problems. *Contraception* **48**(1), 1–12.

Smith G & Pell J (2003) Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ* **327**(7429), 1459–61.

Tanner A, Maiden M & Macuch, P. J. *et al*. (1998) Microbiota of health, gingivitis, and initial periodontitis. *J Clin Periodontol* **25**(2), 85–98.

Temple R (1995) *A regulatory authority's opinion about surrogate endpoints, in Clinical Measurement in Drug Evaluation (eds W. S. Nimmo and G. T. Tucker)*. New York: Wiley.

Theilade E (1987) This week's citation classic. *Current Contents* **1987**(5), 16.

Thompson D (1976) *Uranium in Dental Porcelain*. U.S. Dept. of Health, Education, and Welfare, HEW Publication (FDA) 76–8061.

Thomson W, Broadbent J & Welch, D. *et al*. (2007) Cigarette smoking and periodontal disease among 32-year-olds: a prospective study of a representative birth cohort. *J Clin Periodontol* **34**, 828–34.

Timmerman M, Van der Weijden G & Abbas, F. *et al*. (2000) Untreated periodontal disease in Indonesian adolescents. Longitudinal clinical data and prospective clinical and microbiological risk assessment. *J Clin Periodontol* **27**(12), 932–42.

Tomar S & Asma S 2000 Smoking-attributable periodontitis in the united states: findings from nhanes iii. national health and nutrition examination survey. *J Periodontol* **71**(5), 743–51.

United Nations 2000 *Sources and Effects of Ionizing Radiation*. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2000 Report to the General Assembly.

Valentin J (2005) Low-dose extrapolation of radiation-related cancer risk. *Ann ICRP* **35**(4), 1–140.

Van Damme L, Ramjee G and Alary, M. *et al*. (2002) Effectiveness of col-1492, a nonoxynol-9 vaginal gel, on hiv-1 transmission in female sex workers: a randomised controlled trial. *The Lancet* **360**(9338), 971–7.

Williams R (1990) Periodontal disease. *N Engl J Med* **322**(6), 373–82.

3

The effective use of research data for evidence-based oral health care

Ian Needleman and Helen Worthington

3.1 Introduction

This chapter defines what evidence-based oral health care is and how research data are used in the process. The majority of the chapter will focus on the methodological aspects of generating research data and using data effectively in research synthesis. The chapter mainly considers randomized controlled trials as these are the most thoroughly characterized research design and are considered the highest level of evidence for therapeutic interventions (http://www.cebm.net accessed 28 January 2008). This approach should not be taken to indicate that other research designs, such as epidemiological designs are less important, merely that within a single chapter, it is not possible to treat the topic comprehensively. Assessing the quality of research is essential to its effective use. Therefore, this chapter will also review aspects of bias together with a consideration of standards of reporting research. An introduction to the process of undertaking a systematic review as part of guideline development is also included.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

3.2 What is evidence-based oral health care?

3.2.1 Definitions of EBOH

Evidence-based oral health care (EBOH) is a simple concept but with profound implications for research and for the practice and organization of clinical care. It can be defined as the conscientious, explicit, and judicious integration of (1) best research evidence with (2) clinical expertise, and (3) patient values in order to make decisions about the oral health care of individual patients, groups of patients or populations. Only with integration of the three domains can the outcome truly be described as evidence-based.

3.2.2 Distinguishing between EBOH and evidence

What is often mistakenly termed 'evidence-based' is the production of reviews and guidelines whether employing a rigorous approach to finding, appraising and synthesizing evidence or not. This is more properly termed research synthesis. Whilst research synthesis is an essential component to EBOH it should not be confused with the integrated paradigm that we described above.

In most situations, it is not appropriate for evidence alone to dictate oral health care. There are exceptions however, such as if an intervention was found to cause significant harm and its practice untenable. This is rare in oral health (Moles & Dos Santos Silva, 2000) but was part of the reason for reassessing guidelines for antibiotic prophylaxis for infective endocarditis (see later). More often, evidence needs to be integrated with the patient's preferences and with the skills and experience of the clinician. A consideration of available health care resources is a further component. The reasons for needing integration are:

- The outcome showing a benefit from any particular intervention may not be of interest or value to the patient.
- The value of the benefit may be more statistical than clinical.
- A comprehensive evaluation of the intervention may show benefits in some outcomes and adverse effects from others. Thus the harm-benefit evaluation may be personal to the patient.
- A clinician's judgement will always be essential to determining diagnosis, prognosis and applicability of best evidence.

3.2.3 The promise and limitations of evidence-based oral health

By integrating the patient, clinician and best evidence and by keeping this contemporaneous, the conditions should be most favorable both for fully informed choices and for achieving best health care. Although using evidence in isolation to direct oral health care is not appropriate, readers would be forgiven for believing that

such an approach is acceptable when so many published articles title themselves 'evidence-based'.

Determining what is best evidence is also somewhat crucial and, to this end, various hierarchies of evidence have been proposed (Moles & Dos Santos Silva, 2000; Needleman *et al*., 2005). These hierarchies give more value to some evidence than others (e.g. high quality systematic reviews vs. expert opinion not based on research synthesis). Critically appraising such evidence (see below) is time consuming, difficult to perform and the evidence that teaching these skills improves performance is limited (Parkes *et al*., 2001). Therefore, a major limitation of EBOH is identifying and accessing best evidence. One potential solution is to use only high quality systematic views and guidelines. However, these resources are located in many different databases and can still be difficult to find. Furthermore, the number of such documents is few compared to the number of clinical questions that exist. How to practice EBOH where appropriate research evidence is not available will be more subjective and will require individuals to use the best evidence available (including expert or peer opinion), to extrapolate from other evidence reasonably close to the clinical situation in question and to undertake a risk assessment of the options.

3.3 Why is the effective use of research data important to health care?

3.3.1 Islands in search of continents

It seems self-evident that the effective use of research data is important to health care and to health care research. For instance, using an isolated study to draw conclusions for clinical practice or community health care policy is unlikely to be as effective as knowing and integrating the totality of the evidence. More fundamentally, it is also clear that new research should only be designed with a comprehensive understanding of what is already known. This issue was explored in a paper subtitled 'Islands in search of continents' (Clarke & Chalmers, 1998). The paper detailed an investigation of how well authors of reports of randomized controlled trials (RCT) discussed their results in the context of the totality of the evidence. The findings indicated a problem for health care in that few publications provided a comprehensive consideration of the existing literature in the discussion section. As a result the reader was unable to determine whether the new study presented an isolated finding or whether the results were consistent with other published research. Without this information, the ability of the reader to interpret the results of a single trial and draw conclusions for health care are compromised. Therefore, research data are not being used effectively.

The consequence of not using research data effectively has been demonstrated in a number of publications with striking results. Each has used cumulative meta-analysis as a tool to determine at what point in time enough evidence existed to make clear recommendations about the benefit or harm of an intervention.

Meta-analysis is a statistical tool that allows the results of individual studies to be combined that address a similar question. In a cumulative meta-analysis the studies are combined in chronological order adding an additional study with each new publication. In this way the estimate of how well an intervention is performing can be calculated, together with the precision of the estimate (confidence interval) and the year in which the data achieved statistical and clinical significance deduced.

(Gilbert *et al*., 2005) published a cumulative meta-analysis following a rigorous and comprehensive systematic review of studies investigating infant sleeping position and sudden infant death syndrome (cot death). They found a statistically significant benefit to sleeping on the back by 1970 following publication of two case-control studies (Figure 3.1). However, routine advice to sleep in this position was not made until the early 1990s. The authors estimated that changing the advice in a more timely manner might have prevented over 10,000 infant deaths in the UK and at least 50,000 deaths in Europe, the USA and Australasia.

Whilst observational studies are at greater risk of bias than experimental studies, a similar set of findings as those presented above has been reported for randomized clinical trials. Thrombolytics are drugs used to break up blood clots and are now accepted practice following severe cardiac events such as myocardial infarction. The first clinical trials were published at the start of the 1970s but the drugs

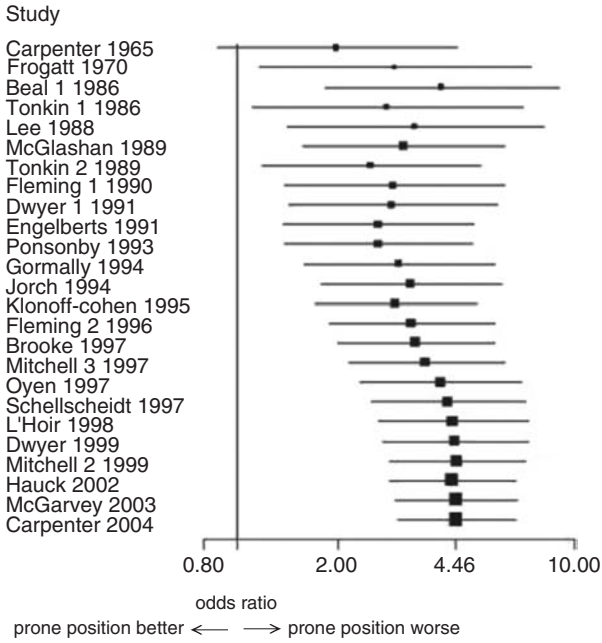


Figure 3.1 Cumulative meta-analysis of mortality from studies investigating sleeping position to prevent sudden infant death syndrome (SIDS) from (Gilbert et al., 2005) with permission.

became accepted into routine clinical practice much later. Predating the analysis of observational studies presented above, an investigation in the early 1990s was performed using cumulative meta-analysis to determine when the evidence for thrombolytic drugs was strong enough to indicate clinical efficacy and then how long it took from this point until routine clinical recommendation (Lau *et al*., 1992). This demonstrated that after eight trials were published in 1973, it was clear that the drugs were effective in reducing death. It also showed that a further 25 trials were published during the next 16 years which involved more than 34,000 individuals, the control groups receiving less effective care. The lag between demonstration of effectiveness and routine recommendation for clinical use was more than a decade. The impact of not receiving best care on the substantial numbers of affected individuals worldwide over this period is clear. Similar findings have been presented for the identification of serious adverse effects with the use of the non-steroidal anti-inflammatory COX-2 inhibitors, the lack of benefit (and possible harm) following the prophylactic use of lidocaine in myocardial infarction and the efficacy for aprotinin in reducing the need for blood transfusion during cardiac surgery.

3.3.2 Case study in oral health: Antibiotic prophylaxis and infective endocarditis

Infective endocarditis (IE) is an infection of the heart lining or valves which continues to have a high rate of mortality. The infection has been associated with bacteria introduced into the circulation during dental procedures. As a result, guidelines from expert bodies have, for the last half-century, recommended the routine use of antibiotics to prevent IE for a wide group of people thought to be 'at risk' of IE. In the last five years, the evidence of effectiveness has been investigated more thoroughly and systematically than previously. As a result, two major national guidelines have made substantial changes to their recommendations, essentially reserving the use of antibiotics to a very small group of patients (Gould *et al*., 2006; Wilson *et al*., 2007). Aspects of effective research usage that brought about this change included:

1. **Research synthesis of evidence of effectiveness** A Cochrane systematic review of the effect of antibiotics for prevention of IE following dental procedures was published in 2008 (Oliver *et al*., 2008). On the expectation that few if any randomized or non-randomized controlled trials would be found, the study searched additionally for cohort and case-control studies where suitably matched control or comparison groups were included. Following an exhaustive search and appraisal of the existing data, the authors found no evidence either to support or refute the use of antibiotics.

2. **Research synthesis of evidence for causation** Bacteraemia following dental procedures is a well recognized phenomenon. It was therefore argued that since the bacteria that could cause IE could be introduced into the circulation during invasive dental procedures, antibiotics should be used to kill the bacteria and prevent onset of IE. However, a comprehensive review of the evidence indicated that the risk of bacteraemia was far

higher following daily impacts such as chewing and tooth brushing (Wilson *et al*., 2007; Roberts, 1999). Therefore, occasional use of antibiotics in individuals previously considered to be at risk made no sense.

As a result of this change, each year, many hundreds of thousands of patients will avoid the use of antibiotics. The evidence for harm of using antibiotics has not been systematically evaluated but estimates suggest that risk of serious side effects such as allergy leading to anaphylaxis (with risk of death) exceeds any possible benefit of prevention.

3.4 Systematic reviews and guidelines for clinical practice

In the previous section, we demonstrated the importance of bringing together the totality of the evidence. There are a number of approaches to achieving this including systematic reviews and clinical guidelines. Clinical guidelines need to assimilate the evidence and they usually use high quality systematic reviews if they exist. Guidelines usually go further and, if systematic reviews or other research data are not available they may reach conclusions based on expert opinion. Following completion of the initial draft of the guideline extensive consultation will follow, inviting comments from major stakeholders including clinicians, patients and, where relevant, industry. Professional associations usually do not have the resources to carry out this type of consultation but they can follow the principles set out in the AGREE protocol which helps guideline writers minimize bias, meet the needs of all stakeholders and maximize clarity (http://www.agreecollaboration.org).

Systematic reviews differ from traditional reviews of the literature in several ways. They are based on and address a focussed question and are undertaken in a systematic manner according to predetermined criteria. These criteria specify which databases are searched, what the inclusion criteria are, how the study quality will be assessed and how the data will be synthesized (Table 3.1). The Cochrane Collaboration Handbook provides great detail about how to undertake systematic reviews of randomized controlled trials (http://www.cochrane.org), and Centre for Reviews and Dissemination at the University of York (http://www.york.ac.uk/inst/crd/index.htm) provides information on the conducting of reviews including study designs other than randomized controlled trials. Systematic reviews are important as they reduce large amounts of information into manageable chunks and as they are used to formulate guidelines and policy they are an efficient use of resources. Systematic reviews may increase the power or precision of the estimate of the relative effectiveness between the interventions being assessed and, if well conducted, should be used to limit bias and improve accuracy.

Meta-analysis is the statistical combination of results from individual studies located by a systematic review. Such an approach is also termed a quantitative synthesis. The potential offered by meta-analysis is that of examining the

Table 3.1 Systematic review process.

Focused question	This includes specifying what the participants are, what the disease/condition of interest is, the interventions being compared and the outcomes to be measured
Study inclusion criteria	Specifying in detail the participants, interventions, outcomes and study design
Search strategy	Which databases are searched, whether experts are being contacted, any hand searching of journals, location of unpublished literature. Any restrictions such as language and the time period covered should be included.
Study validity	Description of the assessment of the validity of each study
Data extraction	Data extraction form piloted and modified. The reliability of the data extracted is important.
Study synthesis	This includes appropriate pooling of the results which may be qualitative (narrative) or quantitative (meta-analysis). The synthesis should reflect the magnitude of the results, the size of studies and the validity of the studies.
Peer review and dissemination to target audience	

consistency of findings between studies addressing the same research question. Meta-analysis also allows the estimation of an overall effect after combining the studies. The statistical combining may overcome inadequate power of small component studies by increasing the precision of the overall estimate, and might allow statistically significant differences to emerge that were hidden in small underpowered studies. With sufficient numbers of included studies, it can also be possible to explore the causes of differences between study results in subgroup analyses. The exploratory analyses might help to derive hypotheses about potential differential effects in the studies caused by for instance different drug doses, stages of disease or impact of methodology and bias. The results of meta-analyses are often presented graphically as forest plots. The forest plots for the meta-analysis from the Cochrane systematic review comparing attachment gain for the periodontal treatment, guided tissue regeneration, compared with the control treatment is shown in Figure 3.2 (Needleman *et al*., 2006). Sixteen trials are included and the mean difference and 95 % confidence intervals are shown for each trial. The trials are in two sub-groups, one of parallel group and one for split-mouth design. A summary estimate is calculated for each subgroup and for all studies combined. Eleven of the 16 trials showed significantly greater mean gain in attachment for the treatment group compared to the control. Overall, the diamond, the pooled

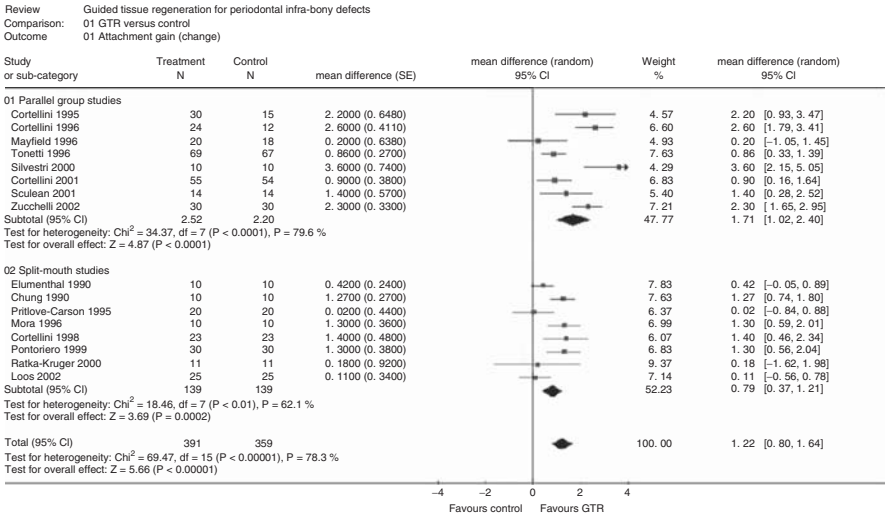


Figure 3.2 Comparisons of guided tissue regeneration versus control for the outcome attachment gain.

weighted summary of the mean differences, indicates the mean difference 1.22 mm and the width of the diamond indicates the 95 % confidence interval from 0.9 to 1.64 mm. The chi-square test for heterogeneity ($p < 0.001$) indicates there is significant heterogeneity which was investigated as part of the review. The statistical methods for undertaking meta-analysis have been developed for continuous, binary and survival data. The statistical software Revman 4 which undertook the meta-analysis and produced the forest plots shown here is freely available from the Cochrane Collaboration website (http://www.cochrane.org). There are some specially developed commercial packages which undertake meta-analysis such as Comprehensive Meta-analysis (http://www.meta-analysis.com/index.html) and MetaWin Version 2.0 (http://www.metawinsoft.com). Mainstream statistical packages such as Stata (http://www.stata.com), SAS (http://www.sas.com) and SPlus (http://www.insightful.com) also include commands to undertake meta-analysis and related statistical procedures.

3.5 Assessing research quality and the effective use of research data

3.5.1 Broad dimensions of health care

The variable quality of research data is a barrier to its use in health care and to incorporation into systematic reviews. However, there are other important barriers that impede the effective use of research data in oral health care. Although this

chapter is mainly concerned with methodological aspects the reader should be aware that data need to be considered in the broadest context of health care. Such aspects will include skills and experience of clinicians, available resources for health care, behavioral change and educational challenges (Niederman & Badovinac, 1999; Monaghan, 1999; Coulter, 2001).

3.5.2 Internal validity

The volume of data in oral health care is vast and increasing daily from approximately 11,000 papers in 1997 to 14,500 in 2006 (search terms 'Dentistry' OR 'Stomatognathic Diseases' in OVID-Medline). The quality of the data will vary between publications suggesting a need to assess this aspect. In this context, we will use quality to mean the methodological quality, also termed internal validity. Internal validity refers to the degree that bias has been minimized in a study (Juni *et al*., 2001).

Bias is any trend in the collection, analysis, interpretation, or publication of data that can lead to conclusions that are systematically different from the truth (Last, 1995). Therefore, the results of a study will be distorted in one direction or other (giving better or worse outcomes) but we will not know either in which direction the bias exists or its magnitude. Consequently, assessment of bias is an essential aspect to using research data. For a RCT, the main forms of bias are shown in Table 3.2.

Selection bias requires a two-step approach to avoidance; firstly the generation of a true random number sequence and secondly, the concealment of the sequence from those responsible for recruiting for the study. Methods to prevent selection bias are listed in Table 3.2.

The impact of improper randomization and concealment on the size of the estimated treatment effect has been investigated in several publications and a systematic review has recently synthesized the evidence (Kunz *et al*., 2007). The systematic review has shown that trials with inadequate randomization or concealment lead to an overestimation of the outcome by 35-40 % compared with adequately randomized trials. A similar finding was reported comparing randomized to non-randomized studies. Therefore, selection bias can have a substantial and measurable impact on the validity of the trial outcome.

Whilst random sequence generation is often reported, methods to conceal allocation are not reported or inadequate. Information from the Cochrane Handbook describing how to assess the quality of these items is presented in Table 3.3. The true test is whether the selection of subjects is entirely unpredictable to recruitment or not. This may not always be obvious. For instance, alternate allocation, odd/even birth dates or hospital numbers have all been used to select subjects for test or control groups. Whilst this might seem random to the trialists, the recruitment will not be unpredictable since those recruiting will know the allocation to the respective experimental group. As a result, there will be an opportunity for recruitment to be biased. Whilst this may be subconscious, reports of deliberate subversion are worrying.

Table 3.2 Principal forms of bias in an RCT.

Bias	Effect	Prevention and comments
Selection bias	The interpretation of a RCT is based on the assumption that populations from which the subjects are drawn are identical with the exception of the test intervention (identical in terms of both known and unknown confounders). Therefore, if there are differences between groups at the end of the study, it is reasonable to conclude that the difference was due to the test intervention. Bias due to distortion of selection might favor allocation of subjects with poorer prognosis to one group biasing the group to a poorer outcome	Generation of true random sequence Concealment of allocation Adequate randomization and concealment is always possible
Performance bias	As with selection bias, if the oral health care provided other than the test intervention, is systematically different between groups, any difference at the end of the study might be due to such a bias	Blinding (masking) of care givers. Achievable with identical placebo controlled studies but may be difficult if not impossible for others.
Measurement bias	Distortion of measurement of outcomes. This is more important where subjectivity (and therefore potential to influence the measurement) is a feature of the outcome. Subjective outcomes include visual assessment of color change and aesthetics or measurement with a periodontal probe (affected by probing pressure). Will be less important for objective outcomes such as tooth loss	Blinding (masking) of examiners Achievable with identical placebo controlled studies but may be difficult if not impossible for others if the intervention produces obvious differences between groups, e.g. comparison of a surgical vs. a drug intervention

(*continued overleaf*)

Table 3.2 (*continued*)

Bias	Effect	Prevention and comments
Attrition bias	Biased losses to follow-up. Even if all the above biases have been controlled, outcomes may be distorted by uneven losses during follow-up. For instance, in relation to benefits, those doing particularly well from the test intervention may fail to return, feeling that their condition is treated. Those that remain will be unrepresentative, in this case towards showing a poorer response. Regarding harms, those that do not return may have experienced unwelcome side-effects and not accounting for this may distort the safety profile	Prevention: Rigorous efforts to maintain adherence of subjects with trial Modeling of impact: Intention to treat analysis Some loss to follow-up is almost inevitable and frequently in the order of 10 % of subjects initially recruited. Intention to treat analyses model conservative assessments of the impact of withdrawals

Performance and measurement bias may be difficult to prevent if the effect of the study intervention is clearly discernible to the subject, caregiver or examiner. For instance, the comparison of a mouth rinse that causes tooth staining such as chlorhexidine will be difficult to mask against another that does not. In such situations, it might be impossible to eliminate bias and the researchers should incorporate a discussion on the potential effects of bias as a limitation to the study. Even where differences are not so clear, masking allocation could be lost. For instance if subjects are not blinded (perhaps due to obvious differences in the appearance of the interventions) they may inform examiners and care givers as to their allocation. For this reason some authors have suggested testing how well masking was maintained such as by asking participants which group they thought they were allocated to (Fergusson *et al*., 2004).

In relation to measurement bias, it has been suggested that objective and subjective outcomes should be considered differently in their susceptibility to bias (Moher *et al*., 1999). For instance, dental implant survival (present or absent) is objective and cannot be affected by the examiner. In contrast implant success criteria are subjective and include measures such as probing depth that may be influenced. Therefore, it is important to have knowledge of the outcome measure in order to evaluate its risk of bias.

Table 3.3 Quality assessment checklist for randomization from Cochrane Handbook.

Item	Classification	Definition
Definition	Adequate	If generated by random number table (computer-generated or not); tossed coin; or shuffled cards
	Unclear	Study refers to randomization but either does not adequately explain the method or no method was reported
	Inadequate	Methods include alternate assignment, hospital number, and odd/even birth date
Allocation concealment	Adequate	Methods include: central randomization (e.g. allocation by a central office unaware of subject characteristics) pre-numbered or coded identical containers which are administered serially to participants on-site computer system combined with allocations kept in a locked unreadable computer file that can be accessed only after the characteristics of an enrolled participant have been entered sequentially numbered, sealed, opaque envelopes
	Unclear	If the study referred to allocation concealment but either did not adequately explain the method or no method was reported.
	Inadequate	Involved methods where randomization could not be concealed, such as alternate assignment, hospital number, and odd/even birth date

Attrition bias may result from patients lost to follow-up or from subjects excluded by the trialists on the basis of violations of protocol. Both of these phenomena may be systematic and not random. For instance, subjects might have taken extra analgesics in a clinical trial of new analgesics due to inadequately controlled pain in one group. Although these subjects might have 'violated' the protocol, excluding their data will bias the outcome. Similarly, subjects who drop-out of a study might be doing so due to systematic differences in their outcomes. All randomized subjects should be included in the analysis and in their allocated group and this is known as an intention to treat (ITT) analysis. If not all data are available then it is customary to refer to analysis of all available participants as an ITT. An 'on-treatment' or 'per protocol analysis' is restricted to subjects who fulfil the protocol and are eligible, and received the intervention. Sometimes studies report both ITT and on-treatment analysis, although the ITT analysis is generally accepted for

Table 3.4 Key aspects of external validity

Domain	Item
Subjects	Are the subjects similar enough to the intended application? Chief characteristics: age, sex, ethnic group, type of disease, severity of disease, presence of risk factors, presence of other oral and systemic diseases
Interventions	Is the intervention provided and assessed in a similar way to the intended application? Skill level, duration, frequency, type of follow-up, appropriateness of timing of outcome assessment in relation to practicability and time course of disease/healing
Settings	Is the setting of the study similar enough to the intended application? Primary vs. other levels of care settings, availability of resources, skill mix of teams

superiority trials there is debate about which is appropriate for non-inferiority or equivalence trials (Brittain & Lin, 2005) and ITT is not recommended for examining adverse effects (Altman *et al*., 2001).

3.5.3 External validity

External validity refers to the generalizability of the results of a study. This aspect will depend on the context in which the results are to be used and will entail more subjective judgement than decisions on internal validity. Nevertheless, assessment of external validity will be as important as methodological issues about bias. The basic question to ask is whether the subjects, interventions and settings are similar enough to the intended application (Higgins & Green, 2008). Judgement will be needed to determine whether small differences are likely to affect health and response to treatment. Table 3.4 summarizes the key aspects to assess.

3.5.4 Critical appraisal checklists

Many critical appraisal checklists are available and can be helpful in learning a structured approach to this difficult skill. Most checklists are specific to a type of health care question and care should be taken to select the appropriate checklist. These can be accessed at the Centre for Health Evidence website (http://www.cche.net/usersguides/main.asp).

3.6 Towards better reporting of clinical research

3.6.1 What are the problems with the quality and reporting of research?

Critical appraisal of research depends on the assumption that publications are reported with enough detail to allow for appraisal. However, a universal finding

Table 3.5 Overview of statistical tests comparing 2 or more than 2 groups for numerical and nominal outcomes. The tests indicated with an asterisk can also be used for ordinal outcomes

Oral health speciality and study	Method	% of trials with adequate reported methods
Periodontology		
Antczak *et al*. (1986)	Randomization	14 %
Montenegro *et al*. (2002)	Randomization	17 %
	Allocation concealment	7 %
Prosthodontics		
Dumbrigue *et al*. (2001)	Randomization	46 %
Jokstad *et al*. (2002)	Randomization	33 %
Implantology		
Esposito *et al*. (2001)	Randomization	16 %
Harrison (2003)	Randomization	3 %

from the medical and dental literature is that this is not the case. Important information is consistently missing from reports and this hinders their assessment. There have been six investigations reported in oral health journals and these include a total of 510 clinical trials across implantology (Esposito *et al*., 2001), periodontology (Antczak *et al*., 1986; Montenegro *et al*., 2002), prosthodontics (Dumbrigue *et al*., 1999; Jokstad *et al*., 2002), and orthodontics (Harrison 2003). The results of these investigations showed that reported randomization methods were adequate in only 3–46 % of studies (Table 3.5) This finding would be of even greater concern if the inadequate reporting reflected actual deficiencies in study conduct. Without universal agreement on standards for reporting, such findings are perhaps not surprising. However, there now exist a number of well recognized standards that have been adopted by most major journals in medicine and several in dentistry.

3.6.2 Standards for reporting of research

The Consolidated Standards of Reporting Trials (CONSORT) statement was first published in 1996 and updated in 2001 (http://www.consort-statement.org). Since then it has gained wide acceptance by most major medical journals with the *British Dental Journal* being the first dental journal to adopt this standard (Needleman 1999). CONSORT comprises two elements; a checklist to guide the author through trial reporting and a flow chart to account for all study participants. Whilst the checklist is intended to be a guide both for the authors and for journal reviewers and not to be published, the flow chart is recommended to be part of the final publication. The items in the checklist are based as far as possible on components with evidence of an effect on study validity. A third version is currently in preparation. The initial guideline was constructed for two group parallel arm studies. Since then, other versions have been adapted for the following trial designs; cluster-randomized, non-inferiority and equivalence and herbal and

medicinal interventions. There is also an extension to strengthen the reporting of harms. Regarding other research designs, reporting standards have been published for systematic reviews (QUOROM, now about to be renamed PRISMA), observational studies (MOOSE) and studies of diagnostic accuracy (STARD). These can all be downloaded from the CONSORT website. A new initiative, the EQUATOR network, was launched in 2008 and aims to bring together guidelines for reporting clinical research in one site and should prove an ideal central resource (http://www.equator-network.org).

3.6.3 Evidence for improvement after adoption of standards

A recent systematic review has investigated whether the quality of reporting does improve following adoption of CONSORT (Plint *et al*., 2006). The findings were that there was better reporting both comparing CONSORT vs. non-CONSORT adopting journals and comparing journals before and after adopting CONSORT. However, the authors also concluded that the evidence was not strong as there was much variability. Indeed, some journals showed a lack of adherence to the guideline following its adoption and others used the obsolete 1996 checklist compared with the current (2001) version.

3.7 Communicating research data

In order to use research data effectively, the data need to be communicated successfully. One substantial challenge for educators, researchers and statisticians is how best to communicate the results of research. It has been shown that the way data are presented can influence its understanding in relation to clinical decision making and this is true both for clinicians and patients (Edwards *et al*., 1999; Gigerenzer and Edwards, 2003). To date, there has been little research investigating this issue in oral health directly although much of the existing development in medical education and informatics should be transferable to the oral health setting. Clearly, communication alone does not mean effective implementation. Achieving implementation of effective oral health care practices is hugely complex and the reader is referred to systematic reviews (Oxman *et al*., 1995) and texts (Clarkson *et al*., 2003) for an introduction to this topic.

3.8 Conclusions

In this chapter we have attempted to give an introduction to the paradigm of evidence-based oral health and to demonstrate how it is based on careful integration of the needs and values of the patient and population, clinical expertise and best evidence. Best evidence depends on the effective use of research data and we have highlighted many of the issues regarding the use of data and aspects of validity and bias. Furthermore, despite the encouraging advances, there continues to be a need to improve the conduct and reporting of clinical research in oral health.

References

Altman D, Schulz K, Moher D, Egger M, Davidoff F & Elbourne D. *et al*. (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* **134**(8), 663–94.

Antczak A, Tang J & Chalmers T (1986) Quality assessment of randomized control trials in dental research. II. Results: periodontal research. *J Periodontal Res* **21**(4), 315–21.

Brittain E & Lin D (2005) A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Statistics in Medicine* **15**, 1–10.

Clarke M & Chalmers I (1998) Discussion sections in reports of controlled trials published in general medical journals: Islands in search of continents?. *JAMA* **280**(3), 280–2.

Clarkson J, Harrison J, Ismail A, Needleman I & Worthington HV (2003) *Evidence Based Dentistry for Effective Practice*. London: Informa Healthcare.

Coulter I (2001) Evidence-based dentistry and health services research: is one possible without the other?. *J Dent Educ* **65**(8), 714–24.

Dumbrigue H, Jones J & Esquivel J (1999) Developing a register for randomized controlled trials in prosthodontics: results of a search from prosthodontic journals published in the United States. *J Prosthet Dent* **82**(6), 699–703.

Dumbrigue H, Jones J & Esquivel J (2001) Control of bias in randomized controlled trials published in prosthodontic journals. *J Prosthet Dent* **86**(6), 592–6.

Edwards A, Elwyn G & Gwyn R (1999) General practice registrar responses to the use of different risk communication tools in simulated consultations: a focus group study. *BMJ* **319**, 749–52.

Esposito M, Coulthard P, Worthington H, Jokstad A, Esposito M & Coulthard, P. *et al*. (2001) Quality assessment of randomized controlled trials of oral implants. *International Journal of Oral & Maxillofacial Implants* **16**(6), 783–92.

Fergusson D, Glass K, Waring D & Shapiro S (2004) Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. *BMJ* **328**, 432.

Gigerenzer G & Edwards A (2003) Simple tools for understanding risks: from innumeracy to insight. *BMJ* **327**, 741–4.

Gilbert R, Salanti G & Harden M (2005) Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002.. *International Journal of Epidemiology* **34**(4), 874–7.

Gould F, Elliott T, Foweraker J, *et al*. (2006) Guidelines for the prevention of endocarditis: report of the Working Party of the British Society for Antimicrobial Chemotherapy. *J Antimicrob Chemother* **57**(6), 1035–42.

Harrison J (2003) Clinical trials in orthodontics II: assessment of the quality of reporting of clinical trials published in three orthodontic journals between 1989 and 1998. *J Orthod* **30**(4), 309–15.

Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008*. Available from www.cochrane-handbook.org <http://www.cochrane-handbook.org/>.

Jokstad A, Esposito M, Coulthard P & Worthington HV (2002) The reporting of randomized controlled trials in prosthodontics. *International Journal of Prosthodontics* **15**(3), 230–42.

Juni P, Altman D & Egger M (2001) Assessing the quality of randomised controlled trials. *In: Systematic Reviews in Health Care. Meta-Analysis in Context*. (M. Egger, S. G. Davey and D. G. Altman (eds.)) London: BMJ Books chapter pp. 87–108.

Kunz R, Vist G & Oxman A (2007) Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews:* Reviews 2007 Issue 2 John Wiley & Sons, Ltd Chichester p. UK DOI: 10 1002/14651858 MR000012 pub2.

Last J (1995) *A Dictionary of Epidemiology*. *Third edition*. Oxford University Press.

Lau J, Antman E, Jimenez-Silva J, Kupelnick B, Mosteller F & Chalmers, T. C. *et al*. (1992) Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine* **327**(4), 248–54.

Moher D, Cook D, Jadad A, Tugwell P, Moher M & Jones, A. *et al*. (1999) Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess* **3**(12), i–98.

Moles D & Dos Santos Silva I (2000) Causes, association and evaluating evidence; can we trust what we read?. *Evidence-Based Dentistry* **2**(1), 75–8.

Monaghan N (1999) Human nature and clinical freedom, barriers to evidence-based practice?. *British Dental Journal* **186**(5), 208–9.

Montenegro R, Needleman I, Moles D & Tonetti M (2002) Quality of RCTs in periodontology–a systematic review. *J Dent Res* **81**(12), 866–70.

Needleman I (1999) CONSORT. *British Dental Journal* **186**, 207.

Needleman I, Moles D & Worthington H (2005) Evidence-based periodontology, systematic reviews and research quality. *Periodontology* 2000 **37**(1), 12–28.

Needleman I, Worthington H, Giedrys-Leeper E & Tucker R (2006) Guided tissue regeneration for periodontal infra-bony defects. *Cochrane Database Syst Rev* (2):CD001724.

Niederman R & Badovinac R (1999) Tradition-based dental care and evidence-based dental care. *J Dent Res* **78**(7), 1288–91.

Oliver R, Roberts GJ, Hooper L & Worthington HV Hooper L Antibiotics for the prophylaxis of bacterial endocarditis in dentistry. Cochrane Database of Systematic Reviews 2008, Issue 4. Art. No.: CD003813. DOI: 10.1002/14651858.CD003813.pub3.

Oxman A, Thomson M, Davis D & Haynes R (1995) No magic bullets: a systematic review of 102 trials of interventions to improve professional practice. *CMAJ* **153**(10), 1423–31.

Parkes J, Hyde C, Deeks J & Milne R (2001) Teaching critical appraisal skills in health care settings. *Cochrane Database of Systematic Reviews:* Reviews 2001 Issue 3 John Wiley & Sons, Ltd Chichester p. UK DOI: 10 1002/14651858 CD001270.

Plint A, Moher D, Morrison A, Schulz K, Altman D & Hill, C. *et al*. (2006) Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust* **185**(5), 263–7.

Roberts G (1999) Dentists are innocent! 'everyday' bacteremia is the real culprit: a review and assessment of the evidence that dental surgical procedures are a principal cause of bacterial endocarditis in children. *Pediatric Cardiology* **20**(5), 317–25.

Wilson W, Taubert K, Gewitz M, Lockhart P, Baddour L & Levison, M. *et al*. (2007) Prevention of Infective Endocarditis. Guidelines From the American Heart Association. A Guideline From the American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee, Council on Cardiovascular Disease in the Young, and the Council on Clinical Cardiology, Council on Cardiovascular Surgery and Anesthesia, and the Quality of Care and Outcomes Research Interdisciplinary Working Group. *J Am Dent Assoc* **138**(6), 739–60.

Part II

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

4

Planning a research project

Timothy A. DeRouen and Donald E. Mercante

4.1 Introduction

The efficient planning of any scientifically valid research project usually involves extensive interactions between a scientific or clinical investigator and a collaborating biostatistician. In carrying out such collaborations, it may be useful to provide structure that will ensure that all pertinent topics are covered in a logical order. In this chapter, we recommend introducing structure to the planning process through the use of our version of '20 questions', denoted Q1 through Q20.

4.2 Defining the research topic

4.2.1 (Q1): What is the question of interest, and how does the question translate into a researchable hypothesis?

The starting point in discussions between an investigator and a biostatistician often sounds like an attempt to communicate between people who speak different languages. For example, the investigator may be interested in knowing if saliva characteristics can be used to help plan the most effective dental treatment, while the biostatistician may be trying to figure out whether testing for the equality of means will somehow apply to this situation. The successful resolution of this problem usually comes about through the two educating each other – the investigator getting more specific about the salivary characteristics that he thinks would be useful in planning dental treatment, and why, and the biostatistician in helping translate those specifics into one or more hypotheses that can be tested. In the saliva example,

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

further discussion may reveal that what the investigator is specifically interested in is knowing whether patients whose saliva has low pH are at increased risk of developing caries, and therefore should be treated with more aggressive prevention strategies. With that understanding, the biostatistician may then be able to help translate that into a researchable statistical hypothesis that there is a significant correlation between a patient's salivary pH level and their past and/or present rate of caries. If the answer to this hypothesis is yes, and there is a significant correlation such that those with low pH levels are at increased risk of caries, then a second hypothesis and related study would be whether such an elevated caries risk in those with low salivary pH could be reduced by a specific intervention.

An acronym that may be useful for investigators to remember in trying to adequately frame a study question, especially if it involves an intervention, is PICO. The 'P' stands for the patient or population of interest; 'I' is for the intervention being evaluated; 'C' is for the control or comparison group; and 'O' is for the outcome used. All four components need to be identified in order to adequately frame a research question (see Chapter 6).

4.2.2 (Q2): What is the outcome of interest?

A key issue in using the PICO terminology to frame a research topic is specifying the 'O' component, or outcome to be used. In studies of life-threatening diseases such as cancer or coronary heart disease, the outcome often of most interest is the ultimate one – death of the patient. In other non-lethal diseases, including most oral diseases (oral cancer being an exception) the disease typically does not cause the death of the patient, and one has to focus on other kinds of outcomes. For some oral diseases (such as periodontal disease) an argument can be made that the ultimate and most relevant outcome is loss of the (affected) tooth. At the same time, arguments can be made that outcomes less extreme and definitive are quite relevant to the patient, such as tooth mobility, pain, ability to chew, and appearance, and could also qualify as outcomes. A key point is to understand the difference between 'true' outcomes and 'surrogate' outcomes. An extensive discussion of surrogate endpoints is provided in Chapter 2.

An important point for the planning process is to identify the primary outcome, the one most important measure, around which the study can be designed. Unfortunately, in many oral health areas such as periodontal disease, there is no consensus on which outcomes are most important and the ones used vary widely (Hujoel & DeRouen, 1995).

4.2.3 (Q3): What is already known about the issue?

In planning any research project, it is important to establish what has already been done, and therefore is known (as well as unknown) about the issue. In seeking funding for the research project, an important component of any application is the background and significance section, which helps establish why it is important to fund and conduct the project (see Chapter 5). Even if external funding is not being

requested for the research project, if it involves humans or animals there will need to be justification for conducting the study. Therefore, it is necessary to succinctly summarize what has been published in the literature on the topic, and use that as a basis for justifying the proposed study. Otherwise, one might be proposing a study that is unnecessarily duplicative of several other previously published studies, or one might be proposing a study for which there is an inadequate scientific basis. Either way, concerns may be raised about ethics or potential waste of resources.

4.2.4 (Q4): What kind of basic, animal, or clinical research information is needed to address the question?

As part of the process of answering the previous question 'What is known about the issue?', one should be able to identify the next logical step in the process. If a principle has been established on a cellular or molecular basis, then the next step may be to show feasibility in a small study in animals. If the results of a small animal study are suggestive but inconclusive, then the next step is likely to be a more definitive (larger) study in animals. If the principle has been established in animal studies, then the next step may be to demonstrate safety and find the optimal exposure (dose) in humans. If the results of animal studies and preliminary (pilot) studies in humans are consistent and conclusive, then the next logical step may be a large, multi-center, randomized clinical trial (RCT). In any case, it is important to identify the kind of information needed in the next logical step in the scientific process.

4.2.5 (Q5): Are there existing databases that can be used to address the question?

For many topics, especially those involving humans, there may be ways of addressing them using existing databases, without collecting new data. In most countries, there are national surveys of the health of a random sample of citizens (e.g. NHANES, the National Health and Nutrition Examination Survey in the U.S. (US National Center for Health Statistics, 2008), China Health and Nutrition Survey (University of North Carolina, Chapel Hill, 2008), NSW Health Surveys of Oral Health in Australia (NSW Health Department, 2000), and in Europe, the European Health for All Database (World Health Organization, 2007)) and numerous health interview surveys have been developed and implemented as a result of the EuroHIS project (Nosikov & Gudex, 2003) that are done on a regular basis, include an oral health component, and are accessible to the public. Also, many US citizens have health insurance that includes dental health. In those cases, the insurance claims processed by a dental insurance carrier contain data that may be a valuable tool in addressing the question (although access to the data would have to be obtained from the insurance carrier). Some dental care may be provided through Health Maintenance Organizations (HMOs) which have large electronic databases of the information on care delivered, along with other oral and medical information. In Canada and many European countries, most of the dental care is provided through government

agencies, which maintain databases of the care delivered. Although most dental care in the US is provided by private practitioners, those patients who qualify for government sponsored Medicaid or Medicare assistance are part of large databases for those programs. In all of these examples, the databases available have shortcomings since they were collected primarily for other purposes, and may not directly provide the quality of scientific evidence that could be obtained from a randomized clinical trial. On the other hand, not every question can be subjected to a randomized clinical trial, and even if it can, the data from these existing databases can provide valuable information in answering the question or refining it further for use in a clinical trial.

4.3 Identifying and developing the best study design

4.3.1 (Q6): If new data must be collected to address the question, what study design is feasible and will produce the highest level of evidence?

The study design of choice should be the one that is feasible and will produce the highest level of evidence. There is a hierarchy of levels of evidence produced by different study designs, which is described in detail for clinical research in Chapters 2 and 3.

Not all study designs are feasible to address a particular question. For example, if one is interested in trying to establish that smoking puts patients at higher risk of developing periodontal disease, a randomized clinical trial in which patients are randomized to smoking or not smoking is not an ethical or feasible design. Therefore, in that situation, one would look for opportunities to conduct cohort studies (either prospective or retrospective) in which self-selected smokers and non-smokers could be compared over time with respect to the development of periodontal disease. If that were not feasible, then one would look at whether a case-control study may be feasible (cases being those who developed periodontal disease, controls being similar to cases but with healthy periodontia, and comparisons between the cases and controls would include smoking history). The key is to select among feasible designs the one that will produce the highest quality of evidence.

4.3.2 (Q7): Given the study design selected, what effect size should the study be designed to detect?

One of the most difficult, and most critical issues that must be addressed in planning a study is the selection of the purported effect size (for a definition, see Chapter 6) that one wishes to detect in the study. In considering effect sizes, several points should be considered:

4.3.2.1 What is the smallest effect that is clinically significant?

In planning many clinical studies, a very important concept is to identify the smallest effect that would be considered clinically significant – i.e the smallest amount

of improvement in outcome that would be considered by clinicians to be important or relevant to clinical practice – and design the study to detect it. However, in some clinical studies the concept of the smallest clinically significant effect is not relevant. For instance in studies where reducing mortality is the objective (whether of individuals or of teeth), how much of a reduction in mortality would we consider not to be of clinical significance?

4.3.2.2 What effect size has been suggested from prior research?

To get an approximate effect size that could be anticipated, it is often useful to search the literature for treatments that could be expected to be similar to the one under investigation. One can then design the study to detect an effect size slightly smaller than those of the published similar treatments. Also, if the literature suggests that similar treatments produced effect sizes that were smaller than the smallest clinically important effect size discussed above, then one might want to rethink the direction of the research, rather than spend time and resources to establish the statistical significance of a treatment effect that will not be considered large enough to be important (see Chapter 20 for an example).

4.3.3 (Q8): For the outcome of interest, what control proportion or amount of inherent variability can be expected?

Once the primary outcome of a study has been identified, it is important to obtain information on how much inherent variability is associated with measurement of a continuous outcome, or what proportion or rate to expect in the control group of a discrete outcome. Determination of the sample size for a study will require a reasonable estimate of these quantities. Inherent variability is the amount a measure will vary from one observation to the next of the same patient/tooth/site (whatever the unit of observation) when nothing has changed. This inherent variability can be due to instrument precision, biologically variation, or a variety of other things that contribute to 'background noise' that keeps repeat observations under the same conditions from being exactly the same. The proportion or rate in the control group is used as a reference value to compare with the experimental response proportion or rate. Estimates of variability or response rates might be obtained from pilot studies designed for that purpose, or from other published studies that used the same outcomes. Whatever the source, the better this estimate the more precise the sample size calculation.

4.3.4 (Q9): What statistical analysis will be used to test the primary hypothesis, and how will the data be presented to convey the results (especially to clinicians)?

Too often, studies are designed and initiated without thinking through the decision-making process of specifying the statistical analysis to be used. Without

adequate biostatistical input, the proposed analysis plan may just state that the means or proportions in the treatment groups 'will be compared statistically'. Only when one has considered the exact hypothesis being tested, and selected the most appropriate statistical method to test that hypothesis, can that information be used to select the sample size needed for the study. Once the most appropriate analysis plan is specified, one should also think about how to present the results in an understandable manner. If the most appropriate statistical analysis to test the primary hypothesis turns out to be, for example, a complex regression analysis, getting clinicians to understand the results (and perhaps change the way they practice) may require presentations in graphical form of simple univariate results. For example, the statistical analysis selected for the Casa Pia children's amalgam trial (DeRouen *et al*., 2002) involved combined Hotelling and extended O'Brien multivariate tests for four primary endpoints. Merely stating that the P-value for a multivariate statistical test was greater or less than 0.05 was not thought (by the journal editor) sufficiently clear to a clinician to be convincing. However, such results augmented with presentations in simplified graphs that demonstrated trends in the data for each of the four primary outcomes were thought to be clearer and more convincing (DeRouen *et al*., 2006).

4.3.5 (Q10): Given the study design, the primary outcome, the postulated effect size, inherent variability or control proportion (rate), along with the statistical analysis plan, what sample size is needed?

Calculating useful sample size estimates is a collaborative effort between the statistician and the research scientist or clinician; see Chapters 2 and 6 for more details on sample size calculations for RCTs. Several methodological approaches exist for calculating sample size. These include methods based on testing a specific hypothesis as described above, or by constructing a confidence interval not to exceed a pre-specified width, or attempting to maximize precision of an estimate while minimizing or controlling cost using Bayesian methods (see Chapter 18). In other situations where complex statistical models are proposed direct computational formula may not exist and sample size estimates can only be obtained through simulation studies involving the postulated model (Lenth, 2001).

In most studies the hypothesis testing approach is used to calculate sample size. A formula based on the specific hypothesis postulated is used to express sample size as a function of the variability or response rates in the outcome variable, the effect size, and the error rates associated with the hypothesis test (see Chapter 10).

After determining the appropriate effect size, the goal in sample size estimation is then to find the minimum sample size necessary to ensure that the specified effect size is detected with high probability (power). Note that in this approach the effect size, estimates of error variability, and postulated error rates are considered known and fixed quantities and the sample size is the unknown to be determined. Often an analyst will produce a table or a graph depicting a range of possible samples sizes obtained by varying the values of the known quantities. This type of analysis

is known as a sensitivity analysis and provides important information as to how sensitive the sample size estimate is to small changes in the inputs.

There are several web sites available which can be easily accessed and used to estimate sample size for different scenarios, including those by developed by Lenth (2008), Schoenfeld (2008), and O'Brien (2008).

4.4 Addressing logistical, funding, ethical, regulatory issues

4.4.1 (Q11): What type and amount of personnel support would be needed to carry out the study?

The skills for estimating the personnel support needed to carry out a study is something that is acquired with experience. If you do not have that experience, then you should look for help from people who do. One of the key decisions is to decide whether the project is large enough to warrant hiring a study coordinator. In most cases, such a person is invaluable and necessary, even if they do not do it full time. An investigator is not likely to have the time, nor the interest, to pay adequate attention to the kind of details necessary for successful conduct of a study, so the first question is likely to be how much of a study coordinator's time is needed, not whether one is needed. Please refer to Chapter 5 for an in-depth discussion of study personnel. In estimating the amount of personnel support needed, some people advocate 'padding' the estimates a bit just to make sure you have included enough support to overcome any oversights. Also, funding sources often will offer to fund a study, but with cuts in the budget submitted. With 'padding', one can make cuts and not significantly hamper the ability to conduct the study.

On the other hand, some will argue that funding sources for oral health research have very limited resources, and they (particularly commercial sources) are not likely to show any interest if padded budget estimates put the costs higher than they are used to paying. Our best advice is to be as accurate as possible, providing details on how the estimates are derived, and be willing to discuss whether your estimates are reasonable. Then, if the best estimates of support personnel needed to conduct the study cannot be covered by the amount offered by the funding source, the amount of shortfall can be intelligently discussed with the funding source and your institution to see how it might be made up. The worst thing you can do is deliberately underestimate the amount of staff support needed to attract financial support, and not have any way to cover the shortfall. The result is likely to be a study of diminished quality.

4.4.2 (Q12): What is the overall estimated cost of the study?

After establishing the amount of personnel support needed for a study, estimating the cost for that support should be relatively straightforward with input from your institutional administrators. Typically, support must also be included to pay for

personnel 'fringe benefits', such as health insurance. Costs for supplies and equipment needed for the study should be itemized and justified. One important issue is whether funds are requested to pay for patient care costs. In some commercially funded studies, an important incentive for patient participation is whether the cost of needed dental care is at least partially covered by the study. In government-funded (e.g. NIH) studies, sometimes the cost of providing needed (ordinary) dental care may not be covered by a research grant, although the added cost for participating in a research project might be. That is the case for studies funded by NIDCR in practice-based networks, but there are exceptions. For example the Casa Pia study on health effects of dental amalgam in children (DeRouen *et al*., 2006) paid for all needed dental care for participants.

In determining the overall cost for conducting a study, many institutions consider the personnel salaries, supplies, and equipment needed for the study as the visible or 'direct' costs, but also require an amount added to those 'direct' costs to pay for the underlying infrastructure or hidden institutional costs for conducting research, usually referred to as 'indirect' or F&A (facilities and administrative) costs. In the US, institutions that receive grants from government agencies negotiate an indirect cost rate with the government which allows it to charge a certain percentage of 'direct' costs (for example, 50 %) as additional 'indirect' costs which become part of the overall cost. The important thing for investigators to understand is that most institutions request, if not require, the inclusion of such indirect costs, so that the total cost the institution requests payment for can be 50–75 % more than the costs estimated for direct project-related costs. If there is a limit on the total costs a funding agency will pay, then the indirect costs have to be included under that limit, or an exception to the institutional policy has to be negotiated. Many dental companies view indirect costs as something they wish to avoid paying, and will insist that they not be charged indirect costs. That can put an investigator in a bind, trying to negotiate between a funding source and his/her institution on some exception to the usual indirect cost charges. There is no simple solution to the problem. Sometimes a compromise can be worked out, sometimes not, and it likely depends on whether the funding source and/or the investigator's institution have any flexibility. An investigator should be aware of the institution's policy before engaging in any discussion with a funding source.

4.4.3 (Q13): What are potential sources of funding for this kind of research?

Sources of funding for research projects usually depend on the nature of the research. If the research deals with efficacy of a new dental material or other commercial product, then the company manufacturing the product is usually the most likely funding source. If the research deals with a more generic issue in oral health, and not some product available only through a commercial firm, then government research funds (such as from NIDCR) may be requested. Government agencies are typically not willing to fund research that will primarily benefit a company. Research funding is also discussed in Chapter 5.

4.4.4 (Q14): What ethical issues involving research on animals or humans are involved, and with whom should they be addressed?

If the study involves either animal or humans, then an investigator should be aware that there are regulations meant to protect the abuse of both in research, and there are institutional committees charged with safeguarding those protections. Training regarding ethical issues in animals and humans is usually required before someone can be a principal investigator on a project involved with either, and institutional committees will need to review and approve a protocol before it can be implemented. For human subjects, the informed consent process is very important, and an Institutional Review Board (IRB) must approve the process and wording of the informed consent document (see Chapter 5). Sometimes the reviews and ultimate approvals by the institutional animal or human subjects committees can require more than one iteration and thereby take several weeks, so investigators should allow adequate time for it. While funding sources may allow applications to go forward prior to getting IRB or animal committee approval, no project will be allowed to be implemented without it.

4.4.5 (Q15): Will the research be submitted to regulatory agencies for review, and if so, will it satisfy their criteria?

The US Food and Drug Administration (FDA) is the federal agency mandated to protect consumer interest by ensuring the safe and effective marketing of drugs and devices in the US. Most dental applications fall in the category of devices, which is regulated by the Center for Devices and Radiological Health (CDRH), one of five regulatory branches of the FDA (US Food and Drug Agency 2008). Section 513 of the Federal Food, Drug and Cosmetic Act (Drugs and Devices) defines three classes of (dental) devices involving progressively greater regulatory control and oversight depending on the level of risk it poses to human use. The Chinese State Food and Drug Administration oversees regulation of pharmaceuticals and likewise defines medical devices into three classes (State Food and Drug Administration 2008). In contrast, the European and Australian regulations define four classes of medical and dental devices, Classes I, IIa, IIb, and III. Directive 93/42/EEC covers the requirements for placing medical devices on the market and putting them into service in Europe (Council of the European Communities 1993). The Australian legislation for medical and dental devices implemented in October 2002 includes the Therapeutic Goods Act 1989 and the Therapeutic Goods (Medical Devices) Regulations of 2002 Australian Therapeutic Goods Administration (2006). The description of device classes follows that conveyed in the US regulations, although classification schemes are similar in other countries.

Class I devices require only general controls such as Good Manufacturing Practice (GMP), accurate branding, appropriate labeling and reporting, and pre-market notification to the FDA (Runner, 2006). Examples of Class I dental devices are hand held instruments, dental drills, and orthodontic appliances.

Class II dental devices require the use of special controls, which may include performance standards, post market surveillance, patient registries, development and dissemination of guidelines. Examples of Class II devices include optical impression systems for CAD/CAM, use of base metal alloy devices, tricalcium phosphate (TCP) granules for dental bone repair, and X-ray machines.

When general and special controls are not sufficient to provide for reasonable assurance of the safety and effectiveness of a device, when it is used in supporting or sustaining human life, presents substantial importance in preventing impairment of human health, or presents a potential unreasonable risk of illness or injury the device will be classified as Class III. Examples include endosseous (intra-bone) implants and bone grafting materials containing a therapeutic biologic (Federal Register, 2005). The CDRH provides a web site useful for classifying any device of interest (US Food and Drug Agency, 2004).

The type of study design and data required for submissions to FDA or other regulatory agencies will depend on its device classification. If the dental product is classified as a drug rather than a device, then submission requirements may change rather dramatically, and consultation with medical colleagues who have been involved in drug trials may be a valuable source of information.

4.5 Attending to details which ensure success

4.5.1 (Q16): How will study participants be recruited/enrolled/retained? Are enrollment goals realistic? If the study is longitudinal, what efforts will be made to ensure adequate retention? Are estimated drop out rates realistic?

Perhaps the most overlooked aspect of clinical research involves recruitment and retention of study subjects. Estimates by clinicians of the number of subjects they can recruit for a study in a given time period, or the amount of time they will need to recruit a specified number of study subjects, are notoriously inaccurate. A good rule of thumb is to ask a clinician for their best estimate of the number of subjects they can recruit for a study, then divide that number by two (or, conversely, ask how long to recruit a fixed number of subjects, then multiply that estimate by two). Investigators should seek advice on, and put resources into, strategies for recruitment. No one approach fits all situations, but the important thing is to realize that recruitment is never as simple and easy as you think. The same is true of retention. It is easy to assume that most subjects who enroll in a study will continue in it. They lose interest, lose motivation, feel they have benefited as much as they can from participation, and for many different reasons drop out. If the percentage of dropouts is too high, then the credibility of the results will be called into question since they might be dramatically different if everyone remained in the study. The point is that it is difficult to overemphasize efforts aimed at recruitment and retention, and effort spent in these areas will pay dividends. At the same time,

caution is suggested in terms of providing incentives to subjects to enroll or stay in a study. There is a fine line between paying subjects a reasonable amount to compensate them for their time and trouble, and paying an amount large enough that it constitutes a bribe. IRBs will usually scrutinize any subject incentive plan to see if it raises ethical issues. Discussion of addition issues related to recruitment can be found in Chapters 5 and 6.

4.5.2 (Q17): Before initiating the study, has a detailed protocol been written that specifies all details needed for conducting the study and collecting the data?

Before a study is launched, particularly if it is a clinical trial, it is crucial that a manual of procedures or operations be written that describes in detail everything that is to be done in the course of conducting the study (see Chapters 6 and 20). Failure to do so will result in 'making it up as you go', which inevitably leads to 'on the spot' decision-making while in crisis mode, which should be avoided. Having a detailed manual of procedures (MOP) also helps minimize problems caused by staff turnover, since a new staff member can go to it to find out what procedure to follow in a given situation. Effort spent writing a detailed MOP prior to launching a study will pay dividends later with fewer problems in conducting the study.

4.5.3 (Q18): Before the study is initiated, are there clear policies on who will be responsible for writing up different aspects of the study for publication, as well as for determining primary and secondary authorship on the publications?

Before carrying out any study, it is always a good idea to develop a publication plan or name a publication committee to identify who will be responsible for writing up various aspects of the study for publication. Misunderstandings about who has responsibility for or will get authorship credit for study publications can ruin the camaraderie essential to an effective collaborative team, whereas discussions in advance can defuse them.

4.5.4 (Q19): Is there a plan for monitoring the study for progress and safety? Does the study require the appointment of an independent Data and Safety Monitoring Board to oversee the progress of the study and monitor for any ethical/safety issues that may arise?

Any clinical trial funded by NIH requires that there be a data monitoring plan for someone to monitor safety issues and assess whether adequate progress is being made. If the study is simple, relatively inexpensive, with little if any risk of adverse effects, the principal investigator may play that role, or someone else

from the same institution may do so. If the trial is large, expensive, and involves some potential risk to participants, then the funding agency will appoint a Data and Safety Monitoring Board (DSMB) (see also Chapter 6) to regularly evaluate the recruitment process, determine if any adverse events are of concern, and assess whether adequate progress is being made to assure that conclusions can be reached at the end of the trial. If the DSMB becomes sufficiently concerned about the occurrence of serious adverse events, a lack of progress, or even that the outcome of the trial can be determined early, they may recommend early termination of the trial.

4.5.5 (Q20): Do you have plans for making the study data accessible to others once the planned analyses have been completed and published?

With the sizable investment government agencies make in studies, there is increasing concern that the data from such studies too often are not fully utilized to produce the maximum amount of information available. Principal investigators often publish the primary results of a study, and perhaps one or two additional papers on secondary outcomes, but many other issues that could be addressed in the data never are. There is increasing pressure (at NIH there is a requirement) that principal investigators develop a plan for making the data accessible to others, once the planned analyses have been completed and published. While this is obviously not the case for proprietary data from studies funded by companies, in the future, in the general spirit of openness in scientific exchange, investigators will likely be expected, if not required, to allow other investigators eventual access to their data sets. It is probably a good idea to think in advance about the timetable and mechanism for doing so.

References

Australian Therapeutic Goods Administration (2006) URL: http://www.tga.gov.au/devices/fs-aeu-class.htm.

Council of the European Communities (1993) *Council Directive 93/42/EEC*. URL: http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri = CELEX:31993L0042:EN:HTML.

DeRouen T, Leroux B, Martin M, *et al*., (2002) Issues in design and analysis of a randomized clinical trial to assess the safety of dental amalgam restorations in children. *Control Clin Trials* **23**(3), 301–20.

DeRouen T, Leroux B, Martin M, *et al*., (2006) Neurobehavioral effects of dental amalgam in children: a randomized clinical trial. *JAMA* **295**(15), 1784–92.

Federal Register (2005) *Rules and Regulations*. URL: http://www.tga.gov.au/devices/fs-aeu-class.htm.

Hujoel P & DeRouen T (1995) A survey of endpoint characteristics in periodontal clinical trials published 1988-1992 and implications for future studies. *J Clin Perio* **22**, 397–407.

Lenth R (2001) Some practical guidelines for effective sample size determination. *The American Statistician* **55**, 187–93.

Lenth R (2008) *JAVA applets for power and sample size*. URL: http://www.cs.uiowa.edu/~rlenth/Power/.

Nosikov A & Gudex C (2003) *EUROHIS: Developing common instruments for health surveys (edited volume)*. IOS Press, Biomedical and Health Research 57.

NSW Health Department (2000) *Report on the 1997 and 1998 NSW Health Surveys*. URL: http://www.health.nsw.gov.au/public-health/nswhs/acknowledgements.htm

O'Brien R (2008) *UnifyPow: a SAS module for sample size analysis*. URL: http://www.bio.ri.ccf.org/Power/.

Runner S (2006) FDA marketing claims, and the practitioner. *J Evid Based Dent Pract* **6**(1), 19–23.

Schoenfeld D (2008) *Statistical considerations for clinical trials and scientific experiments*. URL: http://hedwig.mgh.harvard.edu/sample_size/size.html.

State Food and Drug Administration (2008) URL: http://eng.sfda.gov.cn/eng/.

University of North Carolina, Chapel Hill (2008) *China Health and Nutrition Survey*. The Carolina Population Center, National Institute of Nutrition and Food Safety, and the Chinese Center for Disease Control and Prevention. URL: http://www.cpc.unc.edu/china.

US Food and Drug Agency (2004) *Classify your medical device*. URL: http://www.fda.gov/cdrh/devadvice/313.html.

US Food and Drug Agency (2008) URL: http://www.fda.gov/cdrh/.

US National Center for Health Statistics (2008) *National Health and Nutrition Examination Survey. Datasets and Related Documentation*. URL: http://www.cdc.gov/nchs/about/major/nhanes/datalink.htm.

World Health Organization (2007) *European Health for All Database (HFA-DB)*. URL: http://data.euro.who.int/hfadb/.

5

How to carry out successful clinical studies: lessons from project management

Jocelyne S. Feine, Stephanie D. Wollin and Faahim Rashid

5.1 Introduction

In Chapter 4, important issues in study development and design are discussed. In this chapter, we will discuss specific, practical issues that must be handled throughout the development and implementation of a research study, focusing on randomized clinical trials.

Like any successful business operation, a research study needs good strong management. Common issues that face everyday businesses will also apply to running large-scale clinical trials. The principle investigator needs to understand the importance of building a multidisciplinary team and then managing that team to yield a successful outcome. Basic team management principles such as: holding effective meetings, making your team better than you, setting boundaries, accepting limitations, encouraging people, inspiring loyalty, trusting staff, respecting individual differences, adapting your style to the members of your team, defining expectations, and training staff to bring you solutions instead of problems all apply to the field of scientific research (Templar, 2005). Many of the pitfalls scientists find themselves in could be avoided if they followed pure management principles in their research careers. Understanding management principles and building a strong

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

team will enable the researcher to focus on the science, rather than being bogged down in administrative tasks and troubleshooting.

5.2 Team building

The most important resource necessary for successful clinical research is the research team. Their ability to cooperate, their enthusiasm and their dedication to the study are crucial to its success.

5.2.1 Principal investigator

The principal investigator (PI) should be qualified by education, training and experience to assume responsibility for the proper conduct of the trial. The PI is responsible for protocol development and team assembly. The types of team members that could be involved are: co-investigators, clinician specialists, consultants, study coordinator/monitor, technical assistants, data management person or team and support personnel.

Prior to beginning the study planning process, the PI must confirm that the proposed study will have adequate resources, such as (1) the ability to demonstrate, based on previously gathered data or considered estimates, a potential for recruiting the required number of subjects within the agreed recruitment period, (2) sufficient time to complete the trial within the proposed period, and (3) access to adequate staff and facilities to carry out the trial for the entire study period. Obviously, it is senseless to begin planning a study if these resources are not available.

The PI is responsible for the performance of the study team. Therefore, adequate training of each member of the study team with regards to the protocol, the investigational therapy or product, trial duties and functions must be assured. To increase efficiency, all study-related tasks are divided amongst the team members based on their roles in the study. Most team members participate on a part-time basis; however, the team exists as an organized entity.

5.2.2 Co-investigators

Thought should be put into the choice of those who will be approached to participate in the development, management and orchestration of the study. Recruit researchers or clinician-scientists whose knowledge and skills will contribute to maximize the knowledge potential of the study. Of course, those with the most expertise and experience, as well as those with a successful track record, should be extremely valuable for the study. Of importance is their ability and willingness to communicate and to take responsibility as a collaborative team player. Therefore, you must also consider the collegiality of the person you plan to recruit. A highly experienced and successful researcher may not necessarily be a good team member; personal conflicts, negativity and uncooperative behavior can destroy even the best study plans.

5.2.3 Clinical colleagues as co-investigators

Do not neglect your clinical colleagues with little research experience, because they can be excellent co-investigators by virtue of their intellectual and experiential contribution. Specialists within the field of investigation should be recruited for all trial-related clinical decisions. It is necessary to recruit the number of clinicians required to help treat as many patients as required during the allotted study period. Clinical colleagues should have enough time to devote to your study, and you may need to assist them in communicating this to their administrative heads as they will be taking time from their clinic work to participate in the research.

It is important to meet personally with potential study clinicians to increase the potential for their participation. This will also provide you with the ability to determine whether the clinician is capable of communicating and understanding the rigors inherent in research guidelines. The first rule in successful retention of clinical colleagues is not to hire the wrong investigators to begin with. Some clinicians may not be good candidates because of too strong a belief in one modality or another. At study initiation, clinicians should be in a genuine state of clinical equipoise regarding the merits of each tested therapy in a trial. It is also important that they have not been guilty of misconduct or of non-adherence to protocol in a prior trial.

5.2.4 Project/study coordinator/manager

The project coordinator is essential to a study and is the most important human resource for overall team management, as this person represents the PI in the daily running of the study. This individual should be able to act independently, yet maintain close collaboration with the PI. The study coordinator should be able to work extremely well with others, solicit opinions from the team members, obtain consensus, motivate and encourage team members and allow no room for procrastination. The project coordinator's responsibilities include (1) overseeing subject recruitment; (2) overseeing scheduling of study appointments and procedures; (3) being constantly aware of the status of all ongoing activities within the study related to subjects or clinicians; (4) being the primary resource for developing operational plans; (5) ensuring proper handling of all study related and administrative paperwork, including securing all study data and patient files, ethical renewals, budgetary reconciliation, subject payments etc. In other words, the study coordinator is like the hubcap of a wheel and the centre of all operation aspects, from protocol implementation right through to data management (Figure 5.1). The best project manager is one who thinks ahead and makes plans, someone who can work relatively independently, yet confirm the important issues with the PI. The best project managers are extremely hard to find, so if you get a good one, don't let him/her go!

5.2.5 Biostatistician

Ideally, a biostatistician should be one of the first co-investigators you seek. The advantage to including a biostatistician as a co-investigator is having them

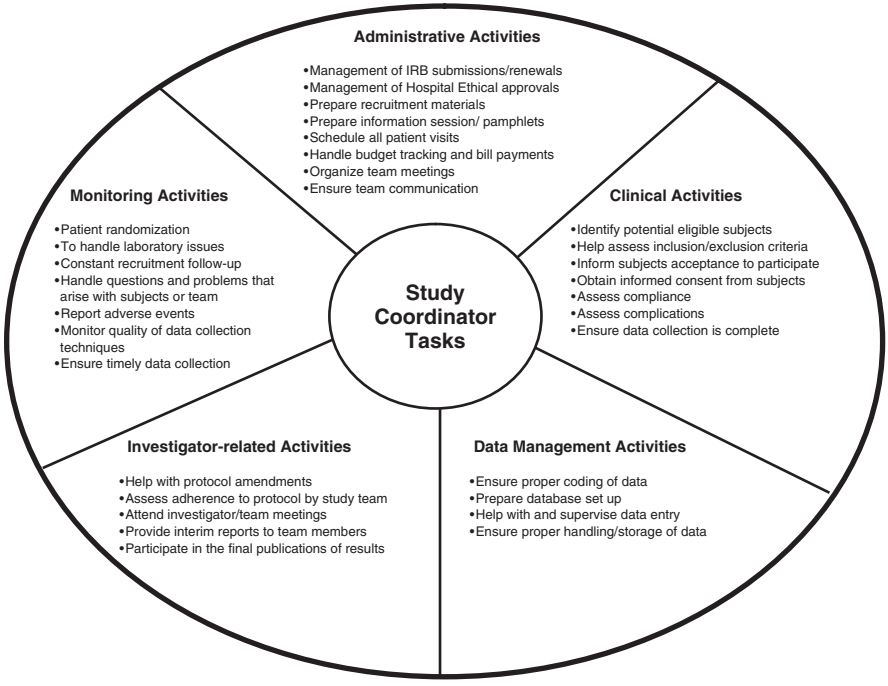


Figure 5.1 Schematic representation of tasks to be performed and/or overseen by the study coordinator/manager.

participate with a vested interest in the research. As a co-investigator, he/she will have a larger role in protocol development and all aspects of data handling procedures from randomization to the final analysis. The amount of time and effort a PI spends working with the biostatistician will increase confidence that the trial is being conducted appropriately. However, if this is not possible, then it is crucial that you hire a biostatistician to consult in the development of the protocol, as well as to prepare the necessary interim statistical reports and to supervise data analyses. Depending upon your resources, the biostatistician may also offer support with data entry and clean-up, as well as the final analyses. The difficulty with hiring a biostatistician is that oral health researchers and statisticians tend to speak different languages. Thus, it may be difficult for the researcher to express exactly what he/she is looking for, and the biostatistician may have limited time to spend, as well as interest in, working on the study. Furthermore, having a statistician as a co-investigator will likely be more cost effective than hiring one as a consultant. In both instances, the best biostatisticians will communicate with you and seek to understand what you are studying, and govern the analysis accordingly. In addition, he/she will take the time to explain what analyses are needed and how the analyses will be conducted (see Chapters 4 and 10).

5.2.6 Clinical research monitor

The purposes of monitoring are to verify that the rights of human subjects are being protected, that the trial data are accurate and complete and that the conduct of the trial is in compliance with the approved protocol. Clinical research monitors require training and should possess scientific or clinical knowledge related to the trial. In addition, the monitors should be familiar with all trial support documents, such as the protocol, informed consent, data acquisition tools and any other written document related to the trial. Trial monitors are sometimes appointed by the trial sponsor (e.g. funding agency, industrial partner). Otherwise, the PI and co-investigators should appoint an individual. These individuals should not be close colleagues or friends of the investigators, i.e. they should have no conflict of interest in order to carry out their duties responsibly. In many smaller scale clinical trials, this job would fall under the umbrella of the 'project co-coordinator', but in large-scale trials in which there are fewer budgetary constraints, another individual could fill the position.

5.2.7 Trainees

Graduate students, post-doctoral fellows and visiting professors may be integral parts of the study team. Graduate trainees must carry out research projects for their programs. Depending on the level of the trainee, your study could be the basis for their work. Depending on the topic, some responsibilities of planning, implementation, data capture, data entry and analysis can be completed by trainees.

5.2.8 Support personnel

The PI must assess if there is a requirement for new personnel or whether personnel who have assisted in previous trials can be used. In order to decide if new personnel need to be hired, all the tasks to be accomplished should be defined and new staff enlisted based on the required skill set necessary for those tasks.

5.2.9 Team management

Prior to study initiation it is the responsibility of the PI to communicate with the research team any plans for publication of the trial findings. All discussions regarding authorship, topics of publication, and involvement of trainees need to be clearly established. It is recommended that these agreements be signed by all participating parties to ensure that there are no ambiguities.

In order to keep the team functioning like a well-oiled machine, open lines of communication are necessary. Regular updates/mailings containing status reports inform team members of the trial's progress. Such communications ensure that all team members are on the same page and make each team player feel important and valued. This is particularly vital when taking on a long-term study because, over time, people tend to become less enthusiastic and require increased

motivation. The morale of the research team is the responsibility of the PI and the study coordinator/manager (Templar, 2005).

5.3 Study initiation

Prior to the initiation of an RCT, it is of great importance that the Principal Investigator and the study coordinator/manager familiarize themselves with 'Good Clinical Practices' (GCP). GCP is an international ethical and scientific standard that is accepted across multiple countries and is expected to be upheld by research teams and institutions worldwide (ICH, 1996).

5.3.1 Obtain funding

As discussed in Chapter 4, the research team needs to work diligently to produce the study budget to be submitted for funding. Careful attention needs to be paid to salaries for support staff, estimated costs of all clinical time, lab work, overhead, and materials and supplies. For support staff, an appropriate percentage for benefits should be included and the salary increased each budgetary year to take into account increases in the cost of living. In order to come up with an accurate representation of what the study will cost, price quotations from any and all participants should be obtained. Industrial partners can provide monetary support, as well as 'in kind' contributions. 'In kind' refers to products that are valued at 40 % of their retail price. Such types of donations can have great effects on the budget of a trial and must be included. In addition, all financial aspects of the trial should be documented in an agreement between the sponsor (e.g. funding agency, industrial partner) and the PI/institution. This document should remain confidential and be kept in a secure place. The financial management of an RCT is a significant task and should be handled by the PI and the study coordinator. Having fewer people involved in fiscal procedures ensures that accurate, detailed, and congruent accounting methods are used.

Budget status and trial expenditures should be tracked from day 1 in order to ensure that the funds will be adequate throughout the entire trial. The use of proper accounting techniques is paramount, as the PI is responsible for reporting the use of all funds and can be subject to audits at any time. Inability to comply with budget stipulations can result in loss of financial support.

5.3.2 Obtain Institutional Review Board approval

One of the documents necessary for every study involving humans is the informed consent. The informed consent is written for potential study participants to explain the nature and procedural aspects of the trial, as well as any risks and benefits related to the study treatments and/or study participation. It should be simple to read and understand. The Institutional Review Board (IRB) of your institution can guide you concerning the sections needed and format that is required.

The steps to obtain IRB approval begin by submitting the final study protocol with all approved amendments, a copy of the informed consent, a copy of all

explanatory information/material to be presented to the subjects and a copy of any advertisement to be used in recruitment. You cannot begin recruitment without Institutional Review Board approval. Further, your institution may not release your research finances without the proper ethical approval of your study. Once the protocol has obtained financial support and IRB approval, the investigator must conduct the trial in accordance with the approved protocol. Any deviations from the protocol must first be submitted to the funders and the IRB for re-approval, except when the changes are logistical or administrative like a new study co-coordinator or phone number changes etc. Should a deviation be made to protect a subject's welfare, then that change must be reported to the IRB and other regulatory bodies as soon as possible; the same procedures apply in the case of an adverse event.

New investigators should expect that this process takes time and that modifications of the consent form, and/or the protocol are often required, which can further slow the process. Therefore, in the interest of time, the investigators may wish to submit the necessary material to their IRB at the same time that they submit the proposal for funding, if not before. If the investigators are unsure about any ethical issue, they can contact their IRB for advice.

5.3.3 Plan study timeline, workflow, and logistics

This involves issues such as estimating and allocating time periods for all aspects of the trial from subject recruitment to publishing reports. Depending on the length of your proposed trial and the nature of various treatment phases the timeline should be geared accordingly. The team must be certain to allocate enough time for the most important phases of a trial, such as recruitment, treatment, data gathering, data entry and data analysis. Please keep in mind that your timeline should even include a final time period for manuscript preparation, submission, and publication.

A study timeline is the same as a 'Gantt' chart in management (Gantt, 1974; Herrmann, 2005). It is developed to give an overview of the predicted periods of time for completion of the study with major time-points indicated. The development of your timeline is an excellent opportunity to consider your resources in terms of your needs, since you must calculate the time necessary to treat participants with the number of available clinicians, office space, support staff, etc. This exercise will also help accurately estimate your costs for budget preparation, so that it is easier to calculate your financial needs from one year to the next. It will be less difficult to prepare the timeline if pilot data on these aspects are available. Without pilot data, it is advisable to be conservative with your predictions. No one has ever been inconvenienced by finishing a study too quickly!

In order to meet the expectations of your timeline and thus complete the trial on time you must organize the workflow of the study in an efficient manor. This means prior to commencing recruitment you must establish logistics, for example:

1. How will you reach your target population?
2. How will prospective subjects contact your team?

3. Will you hold subject information sessions? If so, when and where?
4. What types of information materials will you need to prepare?
5. When and where will you conduct inclusion/exclusion criteria screenings?
6. How many appointments can your clinical staff handle per day?
7. Who will you do your appointment booking?
8. What are your data collection techniques and tools?
9. Are all staff trained to carry out data collection?
10. How and where will the sensitive data be stored and kept secure?
11. What will be the format of your data coding system to protect the identity of the subjects?
12. What is your data entry system? What type of database are you going to use?
13. Who will be handling the statistical analyses? What software will be used? Ensure software compatibility with database extraction format.
14. How much time will be allocated to analyses and interpretation of the results?

The research team's capacity to 'anticipate' potential pitfalls becomes your most valuable tool and, although the above list is not exhaustive, it provides a guideline of the types of seemingly small organizational aspects that need to be addressed prior to beginning your trial. These considerations will greatly improve the efficiency and success of your trial.

5.4 Considerations for subject recruitment

Recruitment practices depend on the study design and protocol. Hospital-based recruitment will be a better way to reach patients with certain health problems or a specific disease state, while outpatient recruitment will be more successful when looking for healthy, community dwelling subjects. Before selecting your methods, the research team must identify the target population.

5.4.1 Outpatient studies

Subject recruitment is different for each type of study undertaken. Some of the ways to reach potential recruits is with Direct–to-Participant Advertising. This includes actions, such as bulletin board notices, media announcements and paid display advertisements in local magazines and newspapers. Such advertisements should include information explaining the nature of the study, the significant

Table 5.1 Cost effectiveness of recruitment advertising (adapted from Perri et al., 2006).

Method	Cost *(can $)*	Number enrolled	Cost/recruited subject (can$)
Major Montreal newspapers	5177.22	25	207.08
Senior newspapers	1917.47	26	73.74
Community newspapers	5932.83	7	847.54
Senior's magazine	6556.47	24	273.18
Public media	0	23	0
Referral	0	20	0
Total	19583.63	127	Mean: 154.20 $

*CAN $ = Canadian dollars

inclusion/exclusion criteria, and whom to contact for more information. Just as with the consent form, all advertisements should be presented in easily understandable common language with minimal use of scientific terminology.

When placing advertisements, the research team should consider carefully where the advertisement is to appear. For instance, are there publications that target the specific population that you are aiming for (i.e. the elderly), or is a more general audience sufficient? The team must also consider which sections and on which days the advertisements should run. Analyzing past recruitment successes can greatly improve a team's selection techniques, saving both time and money (Perri *et al*., 2006). Since finances are always a limiting factor in research studies, the team should address what types of advertisements will bring the most bang for the buck! Keep in mind that the recruitment of subjects is always more costly than anticipated.

For instance, we conducted interim analyses of the cost-effectiveness of our recruitment tools for a clinical trial, during the course of recruitment, in order to eliminate those most costly and least productive (Table 5.1).

Recruitment advertisements will hopefully lead interested subjects to respond. Accurate phone records must be gathered and kept in secure files for future information, such as characterization of the population, success of the advertising medium, etc. In addition, response to potential candidates must be carried out in a timely fashion. The first phone contact can be used as a screening tool to ensure subject eligibility for the trial. This contact with potential subjects is important because this is when the team gets an opportunity for a first impression of a potential study participant. This is vital to determine study compliance. Another important issue to consider during the first contact is that the team presents standard information to each caller so that every potential recruit receives the same information. The team recruiter can then assess willingness to participate. Moreover, this is when potential subjects obtain their first impression of the study team, which makes a big difference in whether he/she decides to participate.

Some additional considerations should be made while recruiting subjects for a clinical trial. Why would a person participate? What potential benefit may someone

get from participating? The best advertising campaign will let patients know that they are satisfying their needs, while helping others at the same time! A promise of personal attention and care by specialists, access to a new treatment and the lure of being part of something significant all play a part in motivating individuals to participate in research studies.

Plan carefully; can you realistically get the number of patients that you are looking for? Are the entry criteria too rigorous? Are you overestimating how much of the population has the particular condition? In order to avoid recruitment pitfalls, continuously monitor the recruitment process and make changes as needed to reach your preset goals. Subject recruitment is a dynamic process that requires creativity, flexibility, patience and a positive attitude.

Follow up on your recruiting efforts to ensure retention of those whom you have enrolled in the trial. Maintain ongoing personal contact with them. Constantly endeavour to show that you recognize the value of the participant's time. There are no clinical trials without willing participants, so careful attention and consideration for participants must be given high priority.

5.4.2 Hospital-based recruitment

One of the first ways to begin hospital-based recruitment is to inform physicians directly about the trial you are conducting. This can be accomplished by holding information sessions for clinical staff at a hospital and expressing a need to reach your desired patient population. Obviously, it is ideal to attract the attention of medical teams that specialize in an area of treatment that meets the research criteria for the trial. Doctors who are willing to be involved may not realize the time commitment necessary to actively recruit subjects for research. Therefore, the PI should plan to allow for extra staff to work directly with the doctor to help complete the recruitment process after the clinician has identified and contacted prospective patients.

Spilker and Cramer (1991) outline the ways to approach physicians about clinical trials. These include the following: (1) The PI should contact the physicians or department chief where the proposed recruitment is to take place. The primary contact can be via letter or telephone, but the PI should make an effort to meet the clinician in person prior to recruitment initiation. (2) The initial contact should provide a brief outline of the study and explanation of the inclusion/exclusion criteria. (3) The PI should explain how recruitment may interfere with the daily routine of the clinician and offer personnel support to assist. (4) The PI should introduce the study coordinator at this time and explain his/her role; the benefits of participation should be explained to the clinician at this time. (5) The clinician and the PI should establish proper channels of communication, i.e. an administrative staff or nurse for the PI or study coordinator to contact as necessary. (6) The hospital unit involved should be left with procedural manuals and reminders so that they can maximize their ability to identify potential subjects from within the hospitalized population. Some hospitals will require that the protocol and informed consent be submitted for additional review by its own ethics committee.

Once you are working with a clinician within a hospital or clinic setting there are a few different ways to reach the population. One very time-consuming method is to carry out chart reviews to identify potential subjects. Each person identified should be confirmed by the attending physician, and the initial contact should be initiated by him/her. Ideally, direct referrals from the doctors themselves will prove more successful, especially since the clinicians are aware of both the study requirements and the patient's health status. In addition, posters can be placed in public areas to attract the attention of potential subjects in the hospital with contact information for the study team.

5.5 Randomization and blinding

Through the premise of unpredictability, the process of randomization produces groups of participants which tend to be alike in terms of known and unknown factors. If carried out as intended, this ensures elimination of investigator bias in the selection of groups, and provides a sounds basis for statistical testing (see Chapter 10).

5.5.1 Importance of randomization

Randomization lists are prepared before the initiation of the study. In smaller studies, procedures are less complicated and may involve the study biostatistician or an investigator who is not directly involved with the study. Randomization in larger multisite studies is carried out at the coordinating centre. The biostatistician may prepare a series of envelopes which hold the sequence of assignment. As patients are enrolled in the study, they receive the next assignment in line. Envelopes are not foolproof and, because of their physical nature, the sequence may easily be distorted. Investigators keen to know the next assignment may tamper with them as well. Multisite studies have used phone-in systems which require investigators to call the coordinating centre where patient eligibility is assessed and patient assigned accordingly (Templar, 2005; Friedman *et al*., 1998). Although this is useful, it prevents the possibility of randomizing after hours or on holidays or in cases where there may be time zone differences. Chapter 6 covers other aspects of randomization.

Interactive voice response systems (IVRS) eliminate the need for human interaction by allowing direct communication between telephone and computer. They have been used in several studies for services including collection of data, intervention delivery and feedback (Abu-Hasaballah *et al*., 2007). Such systems have been developed to handle complex randomization processes, and they make the task easier, less time-consuming and less prone to human error. Eventually, however treatment is assigned; it should be done as close as possible to the point of initial treatment to prevent dropouts after randomization. Such withdrawals should be kept to a minimum. For instance, an increased number of this type of withdrawals prior to treatment could jeopardize reaching the necessary sample size.

Due to ethical constraints, one can only recruit and randomize the number of individuals stipulated within the original protocol. In addition, if you allow too much time to pass between randomization and start of treatment, subjects may change their minds and withdraw, leaving you with fewer participants overall. This is particularly important in trials in which the subjects cannot be blinded to treatment, which is often the case in oral health interventions.

5.5.2 Aspects of blinding

The investigator is responsible for following the designated randomization procedure and should ensure that the randomization code is broken only in accordance with the protocol. If the trial is blinded, the team must inform the sponsor/IRB of any premature unblinding and the reason why this occurred. If possible, blinding (or concealment of allocation) should be implemented because, even if a participating clinician has no strong feelings one way or the other concerning a treatment, he/she may tend to be less conscientious about examining patients he/she knows belong to the control group. He/she may have other unconscious feelings that influence interactions with the patients and, thus, bias the results. Blinding is particularly important where the outcomes being measured are subjective and easily influenced by knowledge of treatment assignment. Blinding makes it difficult for participants of a trial to intentionally or unintentionally influence the outcomes in a trial.

There are various types of blinds, depending on the nature of the intervention and the study design. In an open-label study everyone, including the patient, is aware of the treatment assigned. Single-blinded studies usually keep the patient unaware of his/her treatment, but all others involved have that information. Sometimes, it is the person who enters the data or the statistician who is kept blind, instead of the patient. This would occur when patients cannot be blinded to therapies, such as whether they are getting implant or non-implant prostheses. In a double-blinded study, the patient and the treating clinician are kept unaware of treatment assignment. A full double-blind is when everyone who interacts directly with the patient (research assistant, other staff, etc.) is kept blind. The term ‘triple-blind’ applies to studies in which everyone who has contact with the patients or the investigators is blinded to treatment assignment. In this way, it is hoped that the double-blind is maintained.

Prior to study initiation, methods of blinding should be considered and a guideline should be drawn up in case unblinding occurs accidentally or out of necessity. Factors that could lead to unblinding include adverse drug reactions, efficacy or the lack thereof, changes in lab measures, errors in labeling interventions or even information revealed by unblinded investigators. Each of these needs to be taken into account during the planning stage, and all unblinded team members should be reminded consistently through the course of the trial to maintain discretion.

The validity of the blinding process can be assessed during and after the completion of the trial. Participants who drop out of the study can be asked to guess their assignment. At this point, if they continue to guess correctly, steps should be taken to strengthen the blinding through protocol alteration. End of trial questionnaires

can be used to measure the degree to which participants and investigators/clinicians were blinded. Not only are both parties asked about their possible assignment, they are asked to justify their choices. This is useful to enhance future trials which may involve similar procedures.

Blinded studies are complicated to carry out, and proper care should be taken to ensure accurate coding of interventions after preparation of the randomization list. All packages should be unidentifiable to the delivering clinician and prepared by a third party. Code numbers identifying the treatment assigned are printed on patient forms. Some clinicians may find it necessary to be able to break the blind in cases of patient safety. This should be thoroughly justified, and adequate preparation and communication must be set up between the investigator and the third party to permit this early action.

The maintenance of the blind depends considerably on the type of intervention tested and individual considerations need to be taken. At the end of the day, blinding should only be instituted when there is no harm to the patient and when it is practically feasible to do so.

5.6 Effective management of your data

5.6.1 Data capture

Collect only data that are needed to respond to the study objectives. Collecting extra data can increase cost without any benefit. In addition, collect data based solely on what you plan to analyze. Always use validated and reliable data collection tools, i.e. appropriate biochemical assays, clinician measures and/or, questionnaires. If there is none available for your needs, you must follow specific guidelines in the development of a new and validated data collection tool. Such guidelines can be found within the scientific literature. In general, the chosen collection tool should be measuring what it is intended to measure in a thorough and reproducible manner (see Chapter 16). In addition, a proper data collection procedure manual should be created and implemented in order to decrease inter and intra-observer variation in the data. If special data collection techniques are implemented it may be necessary to offer research team members training in those areas. Studies that require consistency in data gathering should include sessions throughout the period of the study in which the examiners are calibrated and recalibrated.

5.6.2 Data management and security

This aspect of research cannot be treated casually, since many countries now have fixed guidelines concerning data collection and handling regulations. The team must ensure accuracy, completeness, legibility and timeliness of the collected data on a day to day basis. The investigator must maintain trial documents as specified by applicable regulatory requirements; the research team must take measures to prevent accidental or premature destruction of the data documents. The research team must be diligent and abide by the laws of Protection of Personal Information.

In Canada, we have the Privacy Act and the Personal Information Protection and Electronic Data Act (PIPEDA), and the US has the Health Insurance Portability and Accountability Act. These acts exist to protect any information collected that relates to a natural person and allows that person to be identified such as: name, address, phone number, birth date, email address, etc. The participants must sign an informed consent to allow collection of any of the above information. In addition, the study subjects must be informed of the reason the data are being collected, how they will be used and who will have access to them. Further, subjects have the right to know how the information will be protected. The research team must protect against loss, robbery, communication or copy through proper data management techniques.

5.6.3 Data entry

When building your study database, consider programs for data entry. Get and test software to ensure that it meets the requirements of your trial. Ensure availability of hardware for data entry at any and all sites where the trial is being conducted. Personnel should be trained in data input techniques, and the PI should establish if double data entry will be used or whether a spot checking system will be implemented.

Data should be entered in a timely fashion, i.e. as soon as it is gathered, rather than waiting to do so until all data have been collected. Waiting will merely extend the length of the study, needlessly increasing the cost and delaying publication of the results.

Immediate electronic data capture through computer-assisted data entry by the clinician on site (hospital or office) and at the time of the patient's appointment, offers several advantages including:

1. continuous monitoring of data and thus the instant error detection and correction;
2. elimination of other sources of error (no errors in transcription, in copying or recopying);
3. open-ended, easily modified forms;
4. early detection of trends and aides lacking protocol compliance especially vital in multisite studies.

Using immediate computer based methods of entering data provides the opportunity to trial investigators to maintain a vigilant eye on the progress of the trial and allows for deleterious trends to be noticed and dealt with much in advance of trial failure.

5.6.4 Data quality control/assurance

The secret of successful clinical trials lies in maintaining the quality of the collected data. The most frequent sources of error and preventive measures are presented in Table 5.2.

Table 5.2 Potential pitfalls in quality control and possible preventative measures.

Sources of error	Possible preventative measures
Protocol deviations that result when the intervention is not performed/administered as specified	Keep the intervention and the experimental design simple.
Noncompliance of patients with the treatment regimen	Communicate regularly with subjects and monitor progress. Keep the data collected to a minimum, decrease patient burden.
Inaccurate measuring devices	Pretest all questionnaires to detect ambiguities. Use previously validated data collection tools.
Improperly made observations	Prepare a highly detailed procedures manual.
Improperly entered data	Use computer-assisted data entry to catch and correct data entry errors as they are made. Perform frequent audits.

5.6.5 Data management software

With a plethora of clinical trial management software available with varying features, such as Study Manager CTMS® (http://www.clinicalsoftware.net/) and TrialWorks® by ClinPhone (http://www.trialtrac.com/services.aspx), organizing and managing a clinical trial have been made simpler and more efficient. These software programs allow investigators to conduct randomization automatically, capture data electronically, manage individual sites and conduct online tutorials for new investigators. Study protocol information is stored for easy reference and access by all researchers involved in the study.

In addition, subject recruitment and follow-up schedules are maintained to track patient activity throughout the length of the trial. Budgetary planning can be conducted and study financial information can be stored securely. Real time statistical reports are also easily generated. The software can also be used to preserve documentation, such as IRB approvals, as attachments. Best of all, data can be maintained securely through restricted access and encryption and can be accessed quickly, thereby considerably improving the efficiency of the trial and simplifying organization. Data management software programs are key for facilitating smooth clinical trials through easy and continuous progress monitoring. Data management software used through a company will have a built in data extraction application. This provides many options as to what form the data is extracted in. For instance, the team can obtain data in Excel, SPSS and/or SAS formats all at the push of a button. This is very convenient and well worth the investment when it comes time for analysis. Data can also be entered into standard software packages, such as Excel, then converted into biostatistical software programs for analysis.

5.7 Considerations for multicentre trials

Multicentre studies have the potential of producing more meaningful results that are more applicable to the wider population. Larger sample sizes can be collected in shorter periods of time as several centres recruit simultaneously. The use of a variety of practice settings ensures a heterogeneous population with varied sociodemographic characteristics. In addition, multicentre trials also enhance teamwork and allow for a collaborative approach to problem solving between investigators and institutions in the field (Meinert & Tonascia, 1986; Pocock, 1983).

Organization is key to the success of these trials. Multicentre trials are not easy to carry out; they require intense planning and rigorous administration prior to study initiation, particularly because of the distance between sites. Sufficient background information should be available to help in the planning of the trial and to foresee possible design elements that may hinder trial progress. Ideally, multicentre studies need only be carried out when the necessary sample size cannot be recruited at a single site.

Investigator, staff and site selection must be performed carefully. Although a daunting task, proper provisions must be made for communication of ideas through a platform for research teams to share concerns and problems. In situations in which trial design issues are highly debated, the principal investigator must be given the final say, since he/she is the highest level in the chain of command. In order to be constructive, site PIs should offer their opinions freely, yet be open to criticism and the opinions of fellow researchers. Investigators who close themselves off from the rest of the team may weaken the integrity of these trials. From the onset of the trial, the principal investigator must be diligent in selecting site investigators for the study with whom communication is easy and open.

Documentation during the entire trial is critical. All amendments to the study protocol, clinical guidelines and delivery protocols must be noted and circulated to all involved sites. Individual trial protocols may need to be resubmitted to various IRBs for review and acceptance. This is a time-consuming process requiring patience. This may be a more elaborate process if sites are located in different geographic locations. Questionnaires may need to be translated for use at each site. Nothing should be left to memory. Similarly, if investigators debate issues on which they feel strongly, supporting evidence should be made available to prevent unnecessary furor, and no decision should be based on personal opinion. Although several viewpoints may be backed by credible findings, a consensus must be reached among researchers before moving further. These decisions should then be adhered to by all involved.

Data monitoring teams must scrutinize all data received at the coordinating centre to detect and correct errors in a timely manner. The amount of data collected has a direct impact on the quality of the data coming in. Training personnel in data collection can go a long way in ensuring consistency of incoming data. Blunders due to such issues may result in long lasting and catastrophic consequences. Coordinating centres play a vital role in maintaining the scientific integrity of such studies and are responsible to ensure smooth trial flow and conduct.

Another important factor that needs to be dealt with early on in the study planning is the issue of authorship. Investigators who feel that they have been wronged may be less compliant and responsive. In fact, they can block the publication of results. Authorship must be established and agreed upon (ideally in writing) by all investigators/sites early before there is any danger that progress will be hindered (Kraemer, 2000).

5.8 Concluding remarks

As you have read in this chapter, the development and running of a clinical study is multifaceted, requiring patience, attention to detail and excellent management skills. It is not easy because clinical research is complex and challenging. However, the most successful clinical researchers embrace the difficulties and take pleasure in the process.

References

Abu-Hasaballah, K., James, A., & Aseltine Jr, R.H. (2007) Lessons and pitfalls of interactive voice response in medical research. *Contemporary Clinical Trials* **28**: 593–602.

Friedman, L.M., Furberg, C.D., & Demets, D. L. (1998) *Fundamentals of Clinical Trials*. New York: Springer-Verlag.

Gantt, H.L. (1974) *Work, Wages and Profit*. Easton, Pennsylvania: Hive Publishing Co.

Herrmann, J.W. (2005) History of decision-making tools for production scheduling. In: *Multidisciplinary Conference on Scheduling: Theory and Applications*. New York.

ICH (1996) ICH Harmonised Tripartite Guideline. Guideline for Good Clinical Practices. *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*. ICH.

Kraemer, H.C. (2000) Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin* **26**: 533–41.

Meinert, C.L. & Tonascia, S. (1986) *Clinical Trials: Design, Conduct, and Analysis*. New York: Oxford University Press.

Perri, R., Wollin, S., Drolet, N., Mai, S., Awad, M., & Feine, J. (2006) Monitoring recruitment success and cost in a randomized clinical trial. *Eur J Prosthodont Restor Dent* **14**: 126–30.

Pocock, S.J. (1983) *Clinical Trials: A Practical Approach*, Chichester: John Wiley & Sons, Ltd.

Spilker, B. & Cramer, J.A. (1991) *Patient Recruitment in Clinical Trials*. New York: Raven Press.

Templar, R. (2005) *The Rules of Management: An Irreverent Guide for the Leader, Innovator, Diplomat, Politician, Therapist, Warrior, and Saint in Everyone*. Upper River Saddle, N.J.: Pearson/Prentice Hall.

6

Design and analysis of randomized clinical trials in oral health

Brian Leroux and Emmanuel Lesaffre

6.1 Introduction

A randomized clinical trial (RCT) is a prospective study comparing two or more health care interventions in a human population, in which assignments of interventions to trial participants are made at random. The RCT is the strongest study design for evaluation of health care interventions because random assignment of treatment conditions together with blinded assessment of outcomes leads to unbiased estimation of treatment effects. In contrast, observational studies, which rely on self-selected treatment assignments, are susceptible to bias due to confounding between treatment and other factors associated with clinical outcomes. The RCT is the only study design which allows the researcher to infer causal relationships between a risk factor (absence or presence of experimental treatment) and outcome.

Accounts of the history of randomized trials are given in many texts on clinical trials (e.g. Pocock, 1983; Meinert, 1986; Friedman *et al*., 1998). Randomized trials have been widely used in health care starting only in the latter half of the twentieth century. A British Medical Research Council trial of streptomycin for treatment of tuberculosis was one of the early influential studies (Medical Research Council, 1948). Despite the inherent strength of the randomized design, the conclusions drawn from an RCT are only as solid as the integrity of implementation of the study

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

design. The quality of trial implementation varies widely in oral health as in other health fields (see Chapter 1 for some references). Useful guidelines exist to help the researcher design and conduct an RCT that will yield definitive answers, including those of the Cochrane Collaboration (http://www.cochrane.org/) and the US Food and Drug Administration (http://www.fda.gov/). Some clinical trials (particularly those that involve serious health risks to the recruited patients) are conducted under the oversight of an independent Data and Safety Monitoring Board, which is charged with ensuring the integrity of the trial design and implementation, as well as the safety of the trial participants.

This chapter provides an introduction to the basic types of randomized trial designs and to the general principles governing the statistical analysis of trial data. In addition, a summary of the key implementation issues that must be addressed are described. Further details on implementation are provided in Chapters 4 and 5, and in the guidelines referred to above. For details on statistical procedures we refer to Chapter 10 and the subsequent chapters. Further, the website of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH, http://www.ich.org/cache/compo/276-254-1.html) provides numerous recommendations for improvement of the regulatory process of new medicines. The ICH is a unique project that brings together the regulatory authorities of Europe, Japan and the United States and experts from the pharmaceutical industry in the three regions to discuss scientific and technical aspects of product registration. Furthermore, http://www.regsource.com/default.html provides information on regulatory affairs and is also a useful source for those involved in medicinal developments.

First, we discuss what is perhaps the most critical aspect of any RCT, namely, the formulation of the scientific question to be addressed by the trial.

6.2 The scientific question

6.2.1 Phases of clinical research

One of the requirements for a successful randomized trial, indeed for a successful study of any kind, is a well-formulated scientific question. Unfortunately, this has traditionally been one of the most overlooked important features of clinical trial design. In recent years, it has been recognized that the identification of an unambiguous primary scientific question has enormous benefit for all aspects of the trial, from choice of sample size to maintenance of protocol adherence and, ultimately, influence on clinical practice. In addition, a useful distinction has been recognized between the primary question and a set of secondary scientific questions, which are considered less important than the primary question but still within the scope of the trial. By focusing the design and implementation of the trial to address the needs of the primary question, one maximizes the chance of obtaining a definitive answer to it. In contrast, an attempt to answer too many questions, with a lack of focus on any single one, sometimes leads to no answers to any of the questions posed. The choice of primary and secondary questions may depend on the phase of the trial.

Typically for trials involving medicinal products one distinguishes between phase I, II, III and IV studies (Pocock, 1983, pp. 2–3), and sometimes a further subdivision into phase IIIa, IIIb etc. is made (http://www.ich.org/cache/compo/276-254-1.html). However, we briefly describe only the main stages of clinical research.

Clinical research on a new drug typically begins with small studies aimed at determining the acceptability and tolerability of the medication (phase I). Such studies may in fact not be RCTs in the strict sense, not being designed to compare an experimental treatment with a control treatment. Phase I studies also differ from studies in later stages in the choice of study population (tending to use healthy volunteers rather than patients with a certain condition), and in the outcomes considered (focusing on side effects rather than reduction of disease symptoms). Phase II studies are designed to assess the biological activity of the experimental treatment in comparison to a control treatment or a no-treatment control. Such studies use outcomes chosen to maximize the chance of detecting an effect of the experimental treatment in the shortest amount of time; as such, they do not provide answers to the questions of ultimate clinical interest. The phase III trial is designed to give a definitive answer to a question pertaining to clinical practice and is the basis for submitting the experimental treatment to the US Food and Drug Administration (FDA, http://www.fda.gov/) and European Medicines Agency (EMEA, http://www.emea.europa.eu/) for registration. Subsequent to widespread adoption of the experimental treatment, large observational studies are needed to allow detection of rare side effects which cannot be detected in earlier-phase research; these studies are often referred to as phase IV trials, although they are not typically randomized trials.

Clinical trials that involve medical devices such as heart valves in cardiac surgery or tooth filling materials often adopt a different development strategy and are subject to different rules for approval by regulatory agencies. The same is true for trials that evaluate prevention strategies. In the remainder of this chapter all experimental treatments irrespective of their nature (medicinal, medical device, etc.) will be referred to as interventions.

6.2.2 Formulating the scientific question

The general principles of formulating scientific questions for RCTs of all phases are described below using the PICO system (Chapter 4): (1) population, (2) intervention and control, and (3) outcome.

Population A detailed description of the study population is a necessary part of the RCT scientific question. The reason is that in considering a change in clinical practice prompted by the trial, the clinician needs to know to whom the results apply. For instance, a new caries prevention strategy perhaps should not be applied to adults if it is shown to be effective in children. The description of the population includes a detailed list of inclusion criteria which typically specify, at a minimum, the gender, race and ethnicity, severity of disease, and concomitant conditions required of trial participants.

It is inadvisable to overly restrict the set of eligible patients, because that can lead to a trial with little applicability in practice. On the other hand, certain restrictions will be necessary, for example to define the class of patients for whom random treatment assignment is ethical. Note that any generalization of the trial results requires the assumption that the trial participants are representative of the population of interest, which may or may not be valid. For instance, patients who agree to participate in a trial may be inherently different from those who do not agree. The generalizability (i.e. external validity) of the results obtained from a RCT is usually rather limited; often the results of a RCT are generalized to the (wider) population based partly on medical reasoning. Another consideration in defining the study population is the ability to recruit a sufficient number of trial participants. A common mistake in this regard is imposing so many eligibility criteria that patients meeting all the conditions are difficult to identify. Inability to recruit on schedule is the most common reason for failure of a randomized trial.

Intervention and control From the earliest stages of conceptualizing an RCT, the investigator usually has some candidate intervention to be tested. However, it may require considerable effort to sufficiently characterize the intervention. For instance, the frequency and timing of administration of the intervention need to be specified, in addition to the mode of delivery and dose of a medication. Even more difficult in many cases is the choice of the control condition. While it may be of interest to determine the effectiveness of a new intervention compared to a placebo or no treatment at all, the existence of accepted treatments may make it unethical to use a no-treatment control. In such cases, the accepted treatment is used as the (active) control condition. This type of design may lead to a test of equivalence, in which the roles of null and alternative hypotheses are interchanged (see Section 6.5.2). Another issue to be resolved is the number of treatment groups to be compared. For instance, in a phase I study, one might randomize patients to one of several different doses of a drug with the goal of identifying a dose with an optimal trade-off between a desired biological action and unwanted side effects. Another type of multiple group design arises when one wishes to determine the effects of two or more interventions within the same study. The factorial design (see Section 6.3.5) is an efficient way of accomplishing this objective. Although the factorial design can be useful, the interpretation of results can be greatly complicated by the need to consider the possibility of interactions between the treatments. In fact, any design that includes more than two treatment groups is fraught with difficulties in interpretation. For this reason, it is recommended that phase III trials designed to give definitive answers to clinical questions always use two group designs.

Outcome Equally as important as the choice of the population, the intervention, and the control, is the choice of outcome. Often one chooses a primary outcome, which pertains to the primary scientific question, while other (secondary) outcomes may be used in secondary questions aimed at further explaining or supporting the results based on the primary outcome. For a phase III trial the primary outcome variable is preferably an outcome that represents something of tangible benefit or

harm to the patient. These 'true' outcomes (see Chapter 2) are contrasted with 'surrogate' outcomes which represent a measure of a disease process that may or may not have a relation to a tangible patient benefit or harm. For example, patient death is the archetypal true outcome, because it represents a tangible outcome that is readily assessed and of obvious concern to patients. In a classical illustration of the danger of surrogate outcomes (Anonymous, 1989), drugs that were considered beneficial because they eliminated ventricular arrhythmias (the surrogate) were found to increase mortality (the true outcome). Oral health research has traditionally relied on surrogate outcomes, such as pocket depth and caries lesions, although there is recent movement towards the use of true outcomes. For example, tooth-loss has been proposed as a true outcome for trials of periodontal therapies, and measures of oral-health-related-quality-of-life are increasingly being used. Although it is desirable to use a single primary outcome variable in defining the primary question, in some cases it may be impossible to narrow down the list of relevant outcome variables to a single one. In this case, multiple outcome variables may be used, although this practice complicates the design of the trial with regard to the analysis and justification of the sample size. For example, DeRouen *et al*. (2002) used four primary outcome variables in the design of a randomized trial to test the safety of amalgam fillings, which required the development of specialized statistical methods for analysis and power calculations to take into account the multiple outcome variables measured repeatedly over time (Leroux *et al*., 2005).

6.3 Study design

6.3.1 Parallel-arm design

The parallel-arm design is the simplest type of randomized trial. In this design, patients are randomly assigned to one of two or more treatment groups. When the treatment assignment for each patient is made independently of all other patients (for example, using a toss of a coin), this design is sometimes called the completely randomized design to denote the fact that there are no constraints on the random assignments and one patient's assignment does not influence the assignment of another patient. Although this type of randomization is often adequate, there are many different types of randomized trial design that involve constraints on the randomization; these are detailed in the subsequent sections. There is one very simple type of constrained randomization that is often used in the parallel-group design to avoid a large imbalance in sample sizes between treatment groups, which is possible with small studies. By this procedure, the numbers of patients to be assigned to each treatment group are fixed in advance; the identities of the patients assigned to the intervention group are selected using simple random sampling. As an extension of this idea, blocked randomization applies this process to two or more blocks of consecutive patients, thus ensuring near balance in group sizes even if the trial needs to be stopped early.

6.3.2 Stratified (randomized-block) design

Blocked randomization is a special case of the stratified (randomized block) design, in which randomization is constrained to ensure fixed numbers of patients assigned to each treatment group within pre-defined strata ('blocks') of patients. For example, one could create separate lists of random treatment assignments for male and female patients with equal-sized treatment groups within each gender. Note that this implies that the treatment groups are balanced with regard to gender, which is important if there is a strong gender effect on the outcome. Extensions to any number of strata based on 1 or more patient characteristics are possible; however, the number of blocks is limited by the total sample size for the trial.

6.3.3 Crossover (repeated-measures) design

In a crossover design, each patient receives more than one treatment in a randomized order. The simplest case is where half of the patients are randomized to the treatment sequence A-B and the other half to treatment sequence B-A. This design is really just a special type of randomized block in which each patient serves as a block and treatments are randomly assigned to the treatment periods within each block (patient) in a specific but randomly chosen order. Note that A-B refers to a particular sequence in time (first A then B) as opposed to the split-mouth design below, where it refers to an ordering in sites of the mouth. The crossover design is also a special case of a repeated-measures design. The strength of the crossover design is the ability to estimate treatment effects using within-patient comparisons, which may be inherently less variable than the between-patient comparisons used in parallel arm designs because of the avoidance of patient-to-patient variability. However, it is important to recognize the inherent dangers in use of the crossover design due to the possibility of carry-over effects (also called crossover effects), that is, the effect of a treatment which 'carries over' to the next treatment period and influences the apparent response of the patient to other treatments. If different treatments have different carry-over effects (referred to as 'differential carry-over effects') the within-patient treatment effect estimate can be severely biased; in this situation, only the data from the first treatment period for each patient are used to estimate treatment effects and the trial will then be severely underpowered. For this reason, the crossover design is not an option in settings where treatments under study may have permanent effects, e.g. when the patient could be cured (or die) in the first period. If treatment effects are known to be temporary, one can alleviate carry-over effects by using a washout period between two treatment periods to allow the effects of one treatment to be eliminated ('washed-out'). A practical disadvantage of a crossover design is that patients need to stay in the trial for two treatment periods (with a washout period in-between) which increases the likelihood that the patient will drop out prematurely from the trial.

6.3.4 Split-mouth design

Another special case of randomized-block design in oral health is the split-mouth design. In this design, each patient receives all treatments, which are now assigned at random to different parts of the mouth. Thus, one could view a split-mouth design as a kind of crossover design where 'time' is replaced by 'site' in the mouth. As with the crossover design, the ability to perform within-patient treatment comparisons is both a strength and a weakness of the split-mouth design: it increases precision but is susceptible to differential carry-over effects, here called 'carry-across' effects. Because of the high potential for treatment effects to carry over from one part of the mouth to another, this design needs to be used with extreme caution (Hujoel, 1998; Hujoel & DeRouen, 1992; Lesaffre *et al*., 2007).

6.3.5 Designs with two or more treatment factors

Designs for studying two or more treatment factors simultaneously were developed for agricultural research, where for example, researchers may want to study the effect of plant variety and fertilizer type within the same study. In the health field, a researcher may want to examine the effects of a drug (active or placebo) and a behavioural intervention (intense vs. mild) in the same trial. One possibility is to use a factorial design in which patients are randomly assigned to one of four groups: (1) active drug and intense behavioural intervention, (2) active drug and mild intervention, (3) placebo drug and intense intervention, and (4) placebo drug and mild intervention. One of the chief advantages of the factorial design is that it can be used to measure interaction effects, that is the effects due to the treatments 'interacting' with each other. But it must be said that the factorial design is often employed to examine only the main effects of the two factors, assuming (hoping) that there is no interaction at all. In fact, when there is interaction much of the advantage of a factorial design is lost.

The split-plot design offers another way to examine the effect of more than one type of treatment. The name of this design comes from agricultural research and refers to the practice of assigning plots of land to different treatments and then dividing the plots further into 'split-plots' to which levels of a second treatment factor are assigned. In contrast to a factorial design there is now an hierarchy in the allocated treatments. In health research, a split-plot design might involve randomly assigning patients to groups that receive different types of a behavioral intervention and then administering two different drug therapies to each patient in a crossover design. Such designs do not seem to be used in oral health research, probably because of their complexity and the combination of problems related to carry-over effects and difficulties in interpretation related to interaction effects. (Note that the split-mouth design should not be confused with the split-plot design.)

6.3.6 Cluster-randomized design

In oral health research, one often collects outcome data at the level of the tooth or site within a mouth. If the randomization has been performed at the level of the patient (rather than the site as in a split-mouth design), the design may be referred to as a 'cluster-randomized' design to denote the fact that a cluster of teeth or sites is randomized. An analogy may be made to a community-randomized trial, in which entire communities are randomized into one or another treatment group (a useful strategy for testing community-level public health interventions). Specialized methodology has been developed for cluster-randomized designs (Murray, 1998; Donner and Klar, 2000) to account for the correlation between outcomes within a cluster (see also Chapter 13).

6.4 Implementation

6.4.1 Manual of Procedures (MOP)

The creation of a detailed manual of procedures (also called the operations manual) is an essential feature of the proper implementation of a randomized trial design. The MOP describes the detailed steps needed to carry out the procedures of the trial, from identifying potential participants, performing informed consent, enrolling patients, randomizing treatments, administering interventions, entering and analyzing the data. The MOP is an essential resource for training of study personnel and for staff to use a reference when questions arise. Useful guidelines and templates for the development of the MOP are available on the website of the National Institute of Dental and Craniofacial Research (www.nidcr.nih.gov). Typically the MOP would include sections covering all of the aspects of study procedures detailed in the following sections.

6.4.2 Procedures

6.4.2.1 Recruitment

The process by which potential trial participants are identified needs to be explicitly stated at the start. This will be useful to identify the patient population to which the trial results may be generalized. Recruiting from entire intact cohorts, such as student populations of schools or school districts, is appealing because it allows the possibility of generalizing the results to those schools studied and possibly to larger populations of schools if the targeted schools were randomly selected. Of course, generalizability will depend on the recruitment of a large fraction of the potential participants to avoid the possibility of an unrepresentative sample. Successful recruitment requires clear communication of the importance of the trial to the target population to enhance motivation to participate.

6.4.2.2 Informed consent

A critical aspect of trial recruitment is the informed consent process, which is designed to ensure that trial participants are fully aware of the risks and benefits of participation. Procedures for conducting informed consent and the requirements for consent forms and signatures will vary from trial to trial and will need to satisfy regulatory requirements and the approval of a local board responsible for overseeing research involving human subjects (called an Institutional Review Board in the US).

6.4.2.3 Enrolment

It is essential that there be a clearly defined event marking enrolment in the trial to avoid ambiguity about which patients are officially 'enrolled'. In many cases, the event defining enrolment is the assignment of a random treatment assignment to the patient. Alternatively, a patient might be considered enrolled after completing the informed-consent process and giving consent to be a participant in the study. At the time of enrolment, the individual's identity should be recorded in a study participant log and a study identification number assigned to the participant using a predetermined numbering system, including a code for the site in a multisite study as well as a number for the patient. Sometimes, patient initials will also be included in the ID number to help avoid and/or detect errors in transmission of ID numbers, but patient names or other identifying information should not be included.

6.4.2.4 Randomization

The mechanism for conducting the randomization needs to be carefully designed to ensure that it is carried out properly and not subverted in any way. For instance, if a researcher simply created a list of random treatment assignments to be allotted to the patients in the order they are enrolled into the trial, the process could be subverted by trial staff who may choose to enroll or not enroll a patient based on knowledge of the next available treatment assignment. Possible solutions to this problem include the use of sealed opaque envelopes to conceal the random assignments or the use of a computer program to generate random assignments at the time they are needed. In general, randomization should be delayed until it is actually needed by the clinicians administering the treatments. In considering the timing of the randomization it is also important to allow time for the collection of baseline data prior to randomization. It is essential to train study staff in the importance of adhering to the random assignments. It is also necessary to be able to document how the random assignments were produced so that an observer could be satisfied that adequate randomization was used. This requirement usually rules out ad hoc types of randomization such as coin tosses. However, coin tosses may be acceptable if conducted in a public fashion and witnessed by the appropriate individuals; such methods are sometimes used in large community-based trials. Multicentre trials often take place across several different time zones, which makes it difficult to make

available the personnel to actually do the randomization procedure. For such trials an automated (computerized) allocation system is required connected to either the internet or a telephone. For example, the interactive voice response system (IVRS) is a multilanguage automated telephone system that allocates patients to different treatments, which can accommodate stratified randomization. Finally, some trialists prefer minimization over randomization for treatment allocation. Minimization (Scott *et al*., 2002) is a largely non-random adaptive method of treatment allocation with the major aim to minimize the covariate imbalance between treatments.

6.4.2.5 Data collection

Procedures for collecting data need to be described in detail in the manual of procedures. In particular, the study personnel who are to collect the data are to be indicated, the training procedures required for these personnel need to be described, and the plans for subsequent retraining as needed. Plans for assessing reliability of the measures should also be presented if relevant. Final versions of data collection forms, sometimes called case report forms, or CRFs, should be included in the manual. Forms typically undergo changes throughout the course of a trial; hence, it is important to include version numbers on the forms from the start.

6.4.2.6 Treatment administration

The manual should also include a detailed description of the methods for administering the study treatments or interventions. In most trials it is desirable to standardize the treatment administration to a large extent so that the treatments under study can be clearly characterized and the trial results can subsequently be used in clinical practice appropriately. However, in some cases, it may be desirable to allow clinicians to choose some of the details of the treatment delivery according to their usual practices. This has the potential benefit of providing results that are generalizable to 'real-world' clinical settings in which variation between practices exists.

6.4.2.7 Masking

Many trials will involve masking (also called blinding) of treatment assignments to avoid bias due to knowledge of which treatment was delivered. In drug trials, in particular, the different drugs or placebos to be used should be identical in form and appearance to allow as many of the study personnel as possible to be masked. It is especially important that the patient and the study staff member collecting outcome data are unaware of the treatment assignment. On the other hand, the treating clinician may need to be aware of the treatment identity in order to administer it correctly. In general, treatment assignments should be made known only to those who need them for purposes of administering treatments, protecting patient safety, or for maintaining integrity of the trial. At least two individuals should have knowledge of the treatment assignments, for purposes of reporting study results or for unmasking to maintain the patient's safety. One of these would be a data manager or staff member in the coordinating centre. The terms 'single-blind' and

'double-blind' are often used to indicate that the patient only (single-blind) or both the patient and clinician (double-blind) are masked. While double blinded studies are the gold standard procedure, often a single blinded study or even an open trial is all that is possible. For example, when comparing different techniques for tooth restoration, the dentist clearly cannot be blinded although the patient can be (single blinded); when tooth filling materials are compared then even the patient may be able to recognize which tooth has received which material (open trial).

6.4.2.8 Follow-up

Most clinical trials involve follow-up of participants subsequent to treatment administration. This is one of the most critical aspects of the trial implementation for the integrity of the trial. If a substantial percentage (e.g. 20 %) of participants are not successfully followed for collection of outcome data, there is a danger that misleading conclusions could be obtained. The key issue is that participants not followed may differ in important respects from those followed and that the treatment effect estimate is changed by the absence of the lost patients. Statistical methods are available to accommodate missing data to some extent (Chapter 14), which attempt to recover the lost information; however, it is generally not possible to be sure that the information can be reliably recovered. For this reason, it is critical that great efforts are devoted to ensuring that a high proportion of patients are followed up in time.

6.4.3 Data management

Management of the data from an RCT is important to the integrity of the trial. The data management plan must be set out in detail in the manual of procedures in advance of trial initiation, including procedures for collection, transmission, and checking of the data. In multicentre trials, it is essential to have a centralized data coordinating centre to receive and maintain data from all sites to ensure consistency in data management across sites. With current technology it is now possible for data to be entered at clinical sites directly into a server located at the coordinating centre. This system allows timely receipt of data into a system which can include built-in checks on data completeness, validity and consistency. Even with such a system it is necessary to perform additional checks on the data using algorithms for detecting invalid responses or inconsistent responses to two or more questions. The design of these algorithms requires input from both clinical staff and programmers to be most effective. For more details the reader is referred to Chapter 5.

6.4.4 Data and safety monitoring

For ethical reasons, it is essential to monitor the progress and results of an RCT in order to maintain the safety of the participants and ensure that the scientific and ethical rationales for the trial are upheld (Ellenberg *et al*., 2002). For phase III trials in particular monitoring should be performed by an independent group of individuals without ties to the study investigators or their institutions. The monitoring group will meet once prior to initiation of the trial to evaluate the trial design

and plans for implementation in order to ensure that the trial is ethically sound and can be expected to produce scientifically valid data. Subsequent meetings are held annually or more frequently if the participants' potential for risk is high. At these meetings the board monitors the trial results as well as reports of adverse events experienced by the trial participants. A recommendation to stop the trial would be made if it is determined that an acceptable answer to the scientific question has been obtained prior to the planned end of the trial.

6.5 Statistical analysis

In this section we describe the essential statistical issues involved with a RCT. We will omit technical details here. We refer to Chapter 10 for the basic statistical background and to subsequent chapters (or to the published literature) for the more advanced topics.

6.5.1 The statistical analysis plan

The MOP needs to specify the null and alternative hypotheses pertaining to the primary and secondary endpoints. Also, the statistical procedures to follow need to be specified. Besides the MOP another document, called the statistical analysis plan (SAP), is required which contains detailed information on the statistical procedures to follow including those for exploratory analyses. It is important that the SAP is finalized prior to locking the database (after which no changes to the data can be made) to avoid a subjective choice of statistical procedures.

6.5.2 Basic concepts in the statistical analysis of RCTs

Primary and secondary endpoints The simplest design is the two-group parallel group design with only one endpoint evaluated at the end of the treatment period. In that case the statistical analysis is quite simple, involving the most standard statistical tests such as the unpaired t-test for comparing means, the chi-squared test for comparing proportions, and the log-rank test for comparing survival distributions. It is important to realize that the primary and other (secondary, tertiary, exploratory) endpoints fulfill a different role. The primary endpoint is the cornerstone of the RCT (for the study to be called positive, a significant result is needed, say at the 0.05 level of significance, on this endpoint). Secondary endpoints are typically chosen to further support the findings of the primary endpoint and/or to evaluate the safety aspects of the treatments. Tertiary and exploratory endpoints could serve the evaluation of some plausible medical hypotheses to be evaluated in the future. While for a single primary endpoint, no correction for multiple testing (see below and Chapter 10) is needed, such an adjustment may be necessary for the other endpoints.

Superiority and inferiority trials When the aim of the RCT is to demonstrate that the experimental treatment is superior to the control treatment, the trial is

called a superiority trial. All classical statistical tests pertain to testing superiority. On the other hand, when the aim is to show that two treatments are basically equal in performance, then an equivalence trial is needed. Note, however that such a trial can only show that the difference in performance of the two treatments lies in a clinically acceptable range, but not that the two treatments have exactly the same effect. In therapeutic trials, there is much interest to show that the experimental treatment is not (much) inferior to the control treatment. In that case a non-inferiority trial is needed. In the latter two cases the standard classical tests need to be adapted. In Chapter 10 some more information on these types of trials is given. The choice between the three types of designs needs to be made at the planning stage of the trial.

Sample size The sample size calculation (Friedman *et al*., 1998; Meinert, 1986) is an essential part of any RCT in order to minimize the risk of not detecting the aimed effect (if present) of the experimental treatment vis-à-vis the control treatment. That is, the necessary study size needs to be determined to ensure a minimal power (typically 0.80). The sample size calculation varies with the statistical test, is highly technical, and usually requires computer software. Unfortunately, still too many trials are being set up resulting in inconclusive results due to low power.

Intention-to-treat versus per-protocol analyses The choice of which patients the statistical comparison between the two treatments should be based on may seem obvious. However, what should be done when the patient stopped taking medication too early or the patient took the wrong medication or a forbidden concomitant medication? Such deviations from the study protocol give rise to two analysis populations, i.e. sets of patients on which the statistical analysis will be applied. We distinguish the intention-to-treat (ITT) population and the per-protocol (PP) population. The intention-to-treat (ITT) principle is central in RCT research and implies that all patients who have been randomized in the study should be included in the analysis according to the planned treatment irrespective of what happened during the conduct of the trial. This principle has some unexpected implications. For instance, if a patient has been randomized to A but due to a clerical error has been treated with B, then the ITT principle states that for the statistical analysis the patient should be considered as treated with A. Also, if a patient is not compliant with the study protocol procedures, e.g. if (s)he takes prohibited concomitant medication, then that patient still needs to be included in the ITT analysis. Thus the ITT analysis population consists of all patients randomized to the trial medication as intended at baseline. While the ITT principle may seem odd, it is the primary analysis required by FDA/EMEA for registration of a drug in a superiority trial. The reason for this is that the ITT principle is based strictly on randomization, which is the basis for statistical inference in the RCT setting. Note that the ITT principle will often tend to produce a conservative result. Having said this, in practice there are several versions of the ITT principle. Suppose for example that for some randomized patients no values have been obtained for the primary endpoint. These patients cannot be included in an ITT analysis purely

because of lack of measurements, unless missing values are imputed (Chapter 14). Note that in an equivalence or non-inferiority study the ITT analysis is not the primary analysis. Indeed for such a trial the conservative effect of an ITT analysis will be lost and ITT will be biased toward the desired hypothesis (equivalence or non-inferiority) even when the experimental treatment is clearly inferior to the control treatment.

The per-protocol population consists of only patients who have been treated according to the protocol. Thus, a PP analysis excludes noncompliant patients and dropouts and will include patients according to the actual administered treatment. The choice of the PP population involves subjective evaluations of who to include and who to exclude from the analysis and therefore these decisions can only be taken in a blinded manner (blinded to the treatment allocation) and thus after the database lock. For both an equivalence and a non-inferiority trial, regulatory agencies may require that an ITT and a PP analysis are performed and that they show consistent results.

6.5.3 Dealing with the complexity of a RCT

6.5.3.1 Covariate adjustment and random imbalances

Well-planned RCTs should have at least 80 % power to detect the aimed benefit of the experimental treatment. At the planning stage of the RCT it is a challenge to achieve this power with a minimum number of patients. One of the determinants of the necessary sample size is the choice of the primary endpoint. Suppose that the aim of the study is to examine the effect of an experimental antibiotics treatment to cure patients from severe gingivitis problems. Suppose also that a global measure for gingivitis at day 7, denoted as G_7 has been chosen as primary endpoint. One way to lower the study size while maintaining the power is to use increment $G_7 - G_0$ as primary outcome with G_0 the gingivitis measure at baseline. Using the increment instead of the final value of G removes a lot of interindividual variability, thereby reducing the standard deviation of the primary outcome and hence needing a smaller sample size with the same power. A related, but often even more powerful procedure is to use a regression type approach (see Chapter 11) called ANCOVA, whereby G_7 is regressed on the treatment indicator and the value at baseline, G_0. In this procedure G_0 is used as a covariate and the outcome G_7 is adjusted for G_0.

Other covariates such as gender, age, etc could also be included in the ANCOVA model if thought that they will increase the power of the study. Extra covariates could also be included in the regression to remove random imbalance. Randomization implies that the treatment groups are balanced in the population (i.e. same means, proportions, etc.) with respect to all measured and unmeasured patient characteristics but only in the long run. This implies that for any given study size the actual (i.e. achieved) balance is imperfect (one says that there is random imbalance). Covariate adjustment can then, besides increasing the power, also remove the random imbalance and thereby improve the interpretability of the results (Senn, 1989, 1994).

6.5.3.2 Multicentre studies

In medical research, but less so in oral health research, many if not most phase III studies are performed in more than one centre. The main reason for their popularity is that multicentric studies speed up the recruitment of patients considerably. A drawback is that the organization of a multicentric RCT is demanding. The complexity of a multicentre trial starts with the randomization of patients which needs to be done in a stratified manner with centre as stratum (obviously other strata within centres can be taken into account). Further, protocols and hence medical procedures need to be unified across the centres. Many more practical issues need to be resolved. With respect to the statistical analysis, differences with single-centre studies are relatively minor. The statistical tests should allow for the centre effect leading to e.g. Mantel-Haenszel tests for the comparison of a binary outcome, but also to random effects models (see Chapter 13).

In some cases, e.g. when evaluating treatment strategies, surgical interventions, etc., it could be very hard to randomize within a centre. Surgeons could be very reluctant to change their operative technique in a random manner from one patient to another. Also, often a particular treatment regimen is adopted in the whole centre and involves a cascade of personnel; changing this regimen, therefore, in a haphazard way from patient to patient is practically impossible. An alternative approach is then to randomly allocate whole centres to a particular treatment. Such a RCT is a cluster-randomized trial, where the cluster refers to the centre. Typically individuals within centres show some correlation and this should be taken into account in the planning stage (calculating the sample size) but also in the analysis stage of the RCT. In fact there are now two sample sizes that need to be determined: the number of clusters and the number of individuals within clusters.

6.5.3.3 Multiple testing issues

When more than one statistical analysis is needed to decide upon the effect of treatments one should always watch out for the multiple testing problem. Briefly stated the multiple testing problem in a superiority trial consists in an inflated risk (beyond 5 %) of claiming that two treatments are different when they are in fact equal in performance. This can happen in a variety of ways e.g. when there are more than one primary endpoint to be tested, or secondary endpoints that are evaluated. We will consider two important cases in a RCT here: multiple evaluations of the primary endpoint during the conduct of the trial and the analysis of subgroups.

Suppose that for ethical reasons it is necessary to stop the trial as soon as a reasonable answer to the scientific question has been obtained and so one decides to take intermediate looks at the data. To adjust for the multiple views one could apply a Bonferroni correction. This adjustment is, however, too crude for our purposes and therefore dedicated procedures have been developed. When the treatments are evaluated each time the outcome of an individual patient becomes available, one speaks of a sequential procedure. To control the overall Type I error rate to 5 %, the results examined at the intermediate looks should be evaluated with a (much) more stringent significance level each time (much lower than 5 %) in order to

claim the treatments to be different in performance. Pure sequential testing is not very practical. Luckily dedicated procedures have been developed to control the Type I error rate when patients are evaluated in groups or batches. The process of monitoring the results in this way is called (group-) sequential monitoring. A variety of group sequential procedures have been developed such as the Pocock (e.g. Geller & Pocock, 1987) and the O'Brien and Fleming rules (O'Brien and Fleming, 1979). However, they require the specification in advance of the timing for the significance testing. The most popular procedure nowadays is the Lan-DeMets (DeMets & Lan, 1994; Lan & DeMets, 1983) alpha-spending approach which does not strictly require planning the timing of evaluation in advance.

Patients differ in characteristics and hence it is natural to ask the question whether there are subgroups of patients for which the experimental treatment is especially beneficial. Therefore subsequent to a primary analysis often the treatments are compared in a variety of subgroups, e.g. within the group of patients (a) below 65 years of age, (b) above 65 years of age, (c) males, (d) females, etc. Needless to say that the risk of an inflated Type I error rate is greatly increased by such testing. The fact that subgroup analyses are prespecified in the MOP does not alleviate the problem much. While subgroup analyses can be thought provoking they are also not without clinical consequences for future patients. For instance, it could well be that in the total study the experimental treatment proved to be superior to the control treatment but that in none of the subgroups this treatment effect can be established only because the study is not powered for detecting the aimed clinical difference in the (often much) smaller subgroups. Does this therefore mean that there is in fact no beneficial effect of the experimental treatment? As was stated in a debate on the effect of thrombolytic drugs for the treatment of acute myocardial infarct patients, 'subgroup analyses in clinical trials are fun to look at, but one should not believe them' (Sleight, 2000). Indeed, due to a subgroup analysis, the initial conclusion was to restrict thrombolytic therapy to anterior infarction and to deny patients with an inferior infarction access to the thrombolytic therapy. Similarly, due to a subgroup analysis, the initial conclusion was to restrict β-blocker therapy to only anterior infarction patients. Subgroup analyses are also often misused to claim a treatment effect in the absence of an overall significant result. Due to the multiplicity of testing by looking at various subgroups, it is easy to see that by pure chance some results may be statistically significant.

6.5.3.4 Complications during the conduct of the study

Randomization and blinding allow the internal validity (unbiased estimate of difference in treatment effects) of a clinical trial to be high. However, when some of the data is missing and/or patients drop out from the study the internal validity of the study can be greatly affected. The central question here is what can be done when data are missing? A trivial but important observation is that when data are missing they are lost for the analysis. Ignoring the problem of missing data by considering only the patients with available information on the primary outcome is most often a bad choice (see missing-at-random and informative missing data

processes in Chapter 14). A possible way out of the problem is to impute a value for the unobserved data. An imputation technique that was quite popular for many years in RCTs is the last-observation-carried-forward (LOCF) approach. Briefly, this imputation technique imputes the last observed value for the primary outcome for the unobserved value at the end of the treatment period. For example, if the total treatment period is 2 years and every 6 months the primary outcome is measured then, if a patient drops out at year 1, the imputed value for the primary outcome at year 2 is the value observed at year 1. Despite its popularity, the LOCF approach has several deficiencies, i.e. it imputes an unrealistic value for the outcome (not taking into account the natural pattern of the disease and/or of the curing process) and it underestimates the natural variability of the outcome. In the last two decades a lot of research has been done on how to deal with missing data in a proper manner and many approaches have been suggested. The conclusion is that the sensitivity of the results needs to be examined when the missing data are explicitly or implicitly imputed (see Chapter 14).

6.5.3.5 New developments in clinical trial research

Because of economical, but also ethical, concerns there is a constant pressure to speed up the process of clinical trial research. Sequential monitoring allows stopping the trial early when evidence of superiority or harm becomes quickly available. Recently, adaptive study designs have been suggested to move faster from phase II to phase III RCTs and to allow more flexibility when at the planning stage of the trial wrong assumptions or decisions were taken. Important to note is that it is not difficult to switch from one study design to another, the difficulty is doing this while preserving the Type I error rate, guaranteeing unbiased estimates of the difference in treatment effects and maintaining or increasing the power of the study (Chow & Chang, 2006).

6.6 Conclusions

Randomized trials in oral health, as in all fields of medicine, require careful planning and attention to detail to be successful. In addition to the issues described in this chapter, the practical aspects of conducting research described in Chapters 4 and 5 must be carefully addressed prior to initiation of a trial. Dental RCTs often include an additional complexity which arises from the collection of data at the level of a tooth or tooth site, rather than the patient. This entails specialized methods for sample size calculation and statistical analysis (see Chapter 13 for details on analysis of correlated data).

References

Anonymous (1989) Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine* **321**(6): 406–12.

Chow, S-C. & Chang, M. (2006) *Adaptive Design Methods in Clinical Trials*. Chapman & Hall/CRC Biostatistics Series, Boca Rotan.

DeMets, D.L. & Lan, K.K.G. (1994) Interim analysis: The alpha spending function approach. *Statistics in Medicine* **13**: 1341–52.

DeRouen, T.A., Leroux, B.G., Martin, M.D., *et al*. (2002) Issues in design and analysis of a randomized clinical trial to assess the safety of dental amalgam restorations in children. *Controlled Clinical Trials* **23**: 301–20.

Donner, A. & Klar, N. (2000) *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Ellenberg, S.S., Fleming, T.R., & DeMets, D.L. (2002) *Data Monitoring Committees in Clinical Trials: a Practical Perspective*. John Wiley & Sons, Inc., Hoboken.

Friedman, L.M., Furberg, C.D, & DeMets, D.L. (1998) *Fundamentals of Clinical Trials* (3rd edn). Springer-Verlag, New York.

Geller, N.L. & Pocock, S.J. (1987) Interim analyses in randomised clinical trials: ramifications and guidelines for practitioners. *Biometrics* **43**: 213–23.

Hujoel, P.P. (1998) Design and analysis issues in split mouth clinical trials. *Community Dentistry and Oral Epidemiology* **26**: 85–6.

Hujoel, P. & DeRouen, T. (1992) Validity issues in split-mouth trials, *J Clin Periodontol* **19**: 625–7.

Lan, K.K. & DeMets, D.L. (1983) Discrete sequential boundaries for clinical trials. *Biometrika* **70**: 659–63.

Leroux, B.G., Mancl, L.A., & DeRouen, T.A. (2005) Group sequential testing in dental clinical trials with longitudinal data on multiple outcome variables. *Statistical Methods in Medical Research* **14**: 591–602.

Lesaffre, E., Zattera, M.J.G., Redmond, C., Huber, H., & Needleman, I. (2007) Reported methodological quality of split-mouth studies. *Journal of Clinical Periodontology* **34**: 756–61.

Medical Research Council (1948) Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal* **2**: 769–82.

Meinert, C.L. (1986) *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.

Murray, D.M. (1998) *Design and Analysis of Group-Randomized Trials*. Oxford University Press, New York.

O'Brien, P.C. & Fleming, T.R. (1979) A multiple testing procedure for clinical trials. *Biometrics* **35**: 549–56.

Pocock, S.J. (1983) *Clinical Trials: A Practical Approach*. John Wiley & Sons, Ltd, Chichester.

Scott, N.W., McPherson, G.C., Ramsay, C.R. & Campbell, M.K. (2002) The method of minimization for allocation to clinical trials: a review. *Controlled Clinical Trials* **23**: 662–74.

Senn, S.J. (1989) Covariate imbalance and random allocation in clinical trials, *Statistics in Medicine* **8**: 467–75.

Senn, S.J. (1994) Testing for baseline balance in clinical trials, *Statistics in Medicine* **13**: 1715–26.

Sleight, P. (2000) Debate: Subgroup analyses in clinical trials: fun to look at – but don't believe them!, *Current Controlled Trials in Cardiovascular Medicine (now published as Trials)* **1**: 25–27.

7

Epidemiological oral health studies: aspects of design and analysis

Jimmy Steele and Mark Pearce

7.1 Introduction

Epidemiology deals with the distribution and causes of disease in populations, and epidemiological data can do things that other data cannot. They can give an overview, a big picture of the world as it is. At its simplest level epidemiology can give estimates of *prevalence* and *incidence* (see box text in Figure 7.1 – these terms are often used incorrectly). The science of epidemiology can do more than that though. It can identify relationships between risks and populations which can then help to identify potentially causative relationships or pathways, and how they may be modified to improve health. It was epidemiology which initially identified the potential for fluoride to have a role in caries prevention, for example (Dean, Arnold & Elvove, 1942), and many of the now accepted risks for cancer (for example smoking for lung cancer, papilloma virus for cervical cancer) were initially identified by epidemiology and are now acknowledged as causal associations by the public at large. There is sometimes a tendency to over-interpret epidemiological findings, particularly from ecological studies (see below), so it is important to recognize the limitations: epidemiology alone is relatively rarely able to prove cause and effect, particularly from a single study, but it can indicate where and how we might look further.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

Prevalence, incidence and increment

The terms prevalence and incidence are often confused, misused and transposed, but they are completely different concepts. Prevalence is the number of cases of a disease or condition in the population at any point in time, so prevalence can be determined in a cross-sectional survey. Incidence is the number of new cases developing over a period of time, which will usually require either retrospective or longitudinal data. Some conditions are best described by prevalence, others by incidence.

In dentistry, caries can be described using either measure. The proportion of five-year-old children with obvious caries experience in the UK in 2003 was 43 %; in other words, that was the prevalence of obvious caries experience at the age of five years (Pitts *et al*., 2006). The incidence would technically be the number of children who developed caries for the first time over a set time period, usually a year but it depends on the condition and how long it lasts. For example, the incidence of caries might be 7 % per year between the ages of five and six. That would mean that an extra 7 % developed caries for the first time in that year of life. If caries ultimately came to affect everyone in the population then prevalence would be 100 % and subsequent incidence would be zero.

When measuring periodontal disease there are again different sites and teeth, but there are also objective measures of severity for each site, measured by a continuous variable (the number of millimetres of pocketing or attachment loss). This gives accuracy when collecting epidemiological data, but becomes difficult to summarize for a population. Do we describe the number of people with the disease (prevalence), the number of teeth affected (extent) or their maximum pocket depths (severity) and is there any way of combining them? Just to complete the confusion, it is extremely difficult to describe the incidence for periodontal diseases in a meaningful way.

Figure 7.1 Prevalence, incidence and increment.

A good example of the recent use, and potential misuse, of epidemiological data is well illustrated by recent research on the relationship between oral infections (such as periodontal diseases) and coronary heart disease. A whole range of epidemiological study types have been conducted (Scannapieco, Bush & Paju, 2003). Almost all show that before taking into account ‘confounding’ (where two ‘risk factors’ are associated both with each other and the disease or health outcome of interest), people with periodontal disease and other oral sepsis are more likely to have coronary artery disease. This is a statement of statistical fact, but says nothing about the reason for such a relationship or whether it is causal, because both conditions share some of the same risk factors. When these confounding variables, such as smoking, diabetes, social class and many others, are taken into account (statistically), some studies continue to show a relationship and some do not. Even with dozens of studies (good and not so good) conducted over nearly two decades,

many using contemporary techniques to manage the confounding, the nature of any relationship has not really been resolved. This reflects the potential for frustrating imprecision, even with good epidemiology. The lack of certainty about cause and effect is easily overlooked, particularly by those without an understanding of the statistics, with a risk of misuse or misinterpretation of the data. For example, there may be a temptation for policy-makers, the media or even clinicians to use the data to justify decisions that are not supported by the epidemiology. The wholesale removal of diseased teeth might seem justified if there is an association with cardiovascular disease, but there is no evidence at all to support (or refute) such a radical and damaging intervention in terms of cardiovascular outcomes. Nevertheless without epidemiology, the potential for a relationship would not have been identified.

In addition to identifying relationships between conditions, population data are also used for planning health services and social policy. The basics of sampling and analysis are similar, whatever the study or its purpose. The next sections will cover the basic study types and then the principles of sampling populations and ensuring representative data in some detail.

7.2 Study types

There are different kinds of epidemiological studies. Experimental intervention studies, of which clinical trials are the best known example, are often included under the epidemiological banner, but these are sufficiently important in their own right to merit coverage elsewhere in this book (Chapter 6). This chapter is limited to the major types of non-experimental studies as described below.

7.2.1 Ecological studies

Ecological studies look at risks in a geographical or environmental context using aggregated data on entire populations rather than data on individuals. For example, a low mean DMF score may be noted in populations from areas with high fluoride concentrations in the drinking water. The original work on fluoride by Dean (Dean, Arnold & Elvove, 1942) was just such an ecological study and similar approaches are still used to assess the effect of water fluoridation on dental caries in childhood (Riley, Lennon & Ellwood, 1999). Other familiar examples include early work on cancers and environmental risks and more recent work on asthma and infant infections (Pearce, 2000). They are though often prone to a problem known as the *ecological fallacy*. This is where an association seen using aggregated data (e.g. fluoride in the water and low mean population DMF) might not represent the situation at an individual level (e.g. many individual's DMF may be low in an area because that population may just eat less sugar or be from a population which is genetically less susceptible) (Greenland & Robins, 1994). Such arguments are important in the debate about water fluoridation. Nevertheless, well conducted *ecological studies* which take into account such possibilities are very useful if applied appropriately but being descriptive in nature they are often best used for hypothesis generation,

7.2.2 Cross-sectional studies

Cross-sectional studies, such as population surveys, use individual level data at a particular point in time to examine the relationships between health outcomes and potential risks. Such studies can also include the use of retrospective questions to gain some historical information. Cross-sectional studies can also provide an estimate of disease prevalence, but not incidence. Some of the best known examples of cross-sectional studies in terms of dental research are the *national surveys* of oral health that have been undertaken in recent decades in a number of countries including the UK, USA, Ireland, Australia, Finland and several others. In the United Kingdom these cross-sectional studies ran at ten year intervals from 1968–98 (Kelly et al., 2000). Whilst not a *cohort study* (see below), as different people were sampled on each occasion, this repeated cross-sectional approach nevertheless allowed population trends to be plotted for dental indicators as the sampling strategy was the same on each occasion.

7.2.3 Longitudinal and cohort studies

A cohort study refers to a study that follows a group (or cohort) of people over time and can be a powerful way of determining the relationship between an exposure and outcome. The main condition is that no members of the cohort have the disease of interest at the start of the study period. Rates of disease are then compared between exposed and non-exposed individuals. Some cohort studies feature regular follow-ups where additional data, sometimes repeated from earlier stages of the study, are added. These continuing cohort studies are also known as longitudinal studies. They can be difficult to run but have the advantage that with good follow-up they can evolve over time to address contemporary research questions that were never envisaged at the inception of the study. For example, the Newcastle Thousand Families study, a birth cohort established in 1947, has been used to address both tooth retention (Mason et al., 2006) and oral-health-related quality of life (Pearce et al., 2004) in relation to a range of risk factors from different stages of life.

7.2.4 Case-control studies

Case-control studies begin with identification of cases of disease and selection of controls (who do not have the disease). Individual cases are compared with controls, often matched on important confounding or modifying factors, so that the differences in risks between cases and controls can be compared. They are a particularly useful way of studying relatively uncommon conditions because they allow the cases to be selected from the population. This is in contrast to a cohort study where a very large cohort may be required to obtain an adequate number of cases for study. The risk of misinterpreting findings because of poor control selection is quite high. A further form of a case control study is a nested case-control study, where the cases and controls are selected from an existing cohort. This approach is most often used when the measurements to be used in the study are too expensive or time-consuming to be used on the whole cohort.

With the possible exception of ecological studies, these study types are rather common in oral health research. The next section will look at cross-sectional, cohort and case-control studies in more detail and address the more important, practical methodological issues that arise.

7.3 Design issues: cross-sectional studies

Cross-sectional studies seem easy. You take a bunch of people, measure some attributes and then you might calculate the prevalence of a condition or the variation in your sample, or both, allowing you to see how the prevalence is affected by different clinical, social or environmental influences. Sometimes you don't even need to examine anyone, you can just send out a questionnaire to get your data. How difficult can it be? Well, it can be very difficult indeed to do it properly, and if you get the basics wrong some very dangerous conclusions can be drawn. Bad cross-sectional studies are disturbingly easy and disturbingly common and, in our experience, careless sampling is much more of a problem than inappropriate analysis. You can always re-analyse the data from a good sample, but with a bad sample you can do little to resolve the problems with the data you have.

One of the core concepts of all of epidemiology is to do with making sure that the data obtained are from a good sample, in other words they are representative of the population you are studying. Whilst we will discuss them here under cross-sectional studies, many of the statistical principles can be transferred to other study designs to a greater or lesser degree.

7.3.1 Samples and populations

Unless the population you are dealing with is very small, you will normally draw a *random sample* from your population rather than trying to include everybody. For a population numbering any more than a few hundred, surveying everyone would rarely make logistic, economic or even statistical sense. Unless you have a lot of very well organized people looking after your data collection, there is a risk that in trying to do too much you end up with data that are of lower quality (see Chapter 9). The data set needs to contain data from a sufficient number of people to allow you to test your hypotheses or to give sufficient precision to the figures you produce. Your statistical needs (for example the power to test a hypothesis or sufficient precision to estimate prevalence) should therefore determine your sample size, not the size of the starting population.

An everyday example of population sampling is a political opinion poll. In the run up to any election the media is awash with opinion polls. The election itself is when the whole population is surveyed – a rare example of a whole population cross-sectional survey. In an opinion poll only a tiny proportion are sampled, in the UK around 1000 people is normal. The pollsters are always keen to indicate that theirs is a scientifically drawn sample, yet it represents (in the case of the UK) roughly one 50,000th of the electorate. Despite this, the findings are usually

accurate to within a few percentage points. Of course, there are errors, but these are small and the poll generally comes with a declared margin of error (for 1000 people it is around 3 %, based on the confidence limits for the proportion given). If sampling didn't work, no one would want opinion polls and politicians would not hang on their every nuance. That said, opinion poll samples, like samples in oral health research, are subject to distortion for other reasons, particularly response as discussed later. For a statistical view on sampling, see Chapter 10.

As an oral health scientist, drawing a sample will generally make your life easier in any study, but that sample needs to represent the original population from which it was drawn, in just the same way as the opinion poll. In other words, the proportions of different parts of the population need to be accurately represented in your sample. We know, for example, that there are gender differences, age and social gradients in oral health so if the whole population is to be surveyed then any variation between your sample and the population in terms of gender, ages or social class distribution should be random and not systematic (as a result of bad sampling) if the dental data are to be representative. This will give your work 'external validity', so your findings can be generalized to the target population.

7.3.2 Target populations

To draw a representative sample, the *target population* needs to be defined clearly. What is the population of interest? Is it the whole population, or people of specific ages or people suffering from a specific condition? Defining the target population precisely is an important first step.

Usually the population with which we are dealing is finite and a sampling frame can be defined as described below. However, if we are interested in something which might happen, and we need to know where, when or why, our target population may not be finite or easy to define at all. An example may be research on the incidence and risk factors (the variables associated with higher incidence, potentially through a causal pathway) for a rare salivary tumour. In this case there is no obvious part of the population to start with. If a sampling frame cannot be defined, alternative approaches to ascertaining such data may be required, such as reported cases, hospital admissions, mortality data and so on. This will result in a different kind of data set which will not necessarily be statistically representative of any given subset of the population but can still provide important cross-sectional data. This is often the approach used for case-control studies, as described later.

7.3.3 Sampling frames

Assuming that a target population is finite and can be defined, a *sampling frame* is required from which a proportional sample can be drawn. The sampling frame is the list, or nowadays usually the electronic database, which contains the names of that target population. Assuming that your sampling frame is representative in the first place, and assuming that your sample is drawn by random selection of individuals from the frame, there will be no systematic sampling error which may lead to bias.

For example, the proportion of men and women, or affluent and deprived people should be similar to that in the original sampling frame and any deviation should be small and due to chance. Random selection is absolutely critical.

The key feature of the sampling frame is that it should contain as complete a set of names or details from your target population as possible. Where the sampling frame is incomplete, there is a risk that missing data distort the sample before you even start. Getting the best sampling frame for your purposes is not always easy, not least because of contemporary data protection and governance issues. Examples of sampling frames used in dental surveys of large populations have included:

- Postal or zip codes, often using different levels of code for the random sampling process. This is good but may involve a lot of footwork as you usually need a researcher to go to the addresses.
- Government or local authority lists of their populations, for example electoral rolls. These are good but many people are not on the electoral roll, particularly those from low socio-economic backgrounds, so there may be potential biases that could be related to oral health.
- The telephone directory. Straightforward, but misses anyone without a telephone line and anyone who does but has asked not to be in the directory, again with potential biases. The widespread use of mobile phones means this is probably now a greater problem.
- Lists from medical practices. This requires that almost everyone in the population is registered with a family doctor. Some subgroups may be under-represented, particularly where the healthcare system means that primary care is not free at the point of delivery.
- For a survey of professionals, the list of registered professionals held by a national or regional registration body.

These are all examples of appropriate sampling frames, despite the problems identified. There are also many examples of bad sampling frames. For example, a list of patients attending a *dental practice* may make a reasonable sampling frame for the dental patients attending that particular practice, but would be a very poor sampling frame for a survey of the population's oral health because that practice will not be representative of the population, notwithstanding the fact that it will miss dental non-attenders completely. Consequently, a single dental practice is not a great place to sample for population studies.

However, there may be times when dental practices are perfectly legitimate locations for epidemiological research, for example if we are interested specifically in regularly attending patients. Data from one practice would not be generalizable, but it is possible to sample randomly, taking several practices from across a country or region as the first sampling unit, perhaps then followed by a random sample from within each practice. That would allow a regionally or nationally representative sample of dental patients to be drawn. As always, great care and attention to

detail is required, but it is possible. Such *multistage sampling* strategies are used widely as they can make data collection much more cost effective than a single stage strategy. In the UK national surveys of adult oral health, postcode units were randomly sampled first and then households randomly sampled from within these postcode units. It means that researchers can gather data in multiple compact geographical areas, reducing travel time and cost.

The easiest sample is the *'convenience' sample*, usually meaning 'from no sampling frame at all'. Convenience samples, representing whoever you can get to take part, are sometimes a reasonable basis for case control samples, and are often a perfectly legitimate approach to sampling for clinical trials (see Chapter 6), but are not a safe and accurate way to survey the population (Chapter 10).

Of course, a good sampling frame does not automatically mean a good sample will be drawn. The sample itself will need to be drawn randomly, it needs to be of sufficient size to allow for errors in the frame (double entries, clerical errors, people who have moved or died) and most of all, there needs to be a reasonable *response rate*.

7.3.4 Response rates

Even with a perfect sampling frame, ethical considerations normally mean that the members of the randomly selected sample have a choice about whether or not to take part in the study. The opinion poll is a good example once again. When the pollster rings the randomly selected telephone number, the target person may answer and take part, answer and refuse to take part, or simply not answer. Each may mean something different; refusal and *non-response* mean different things in terms of how representative the final sample is. The concern with both refusal and non-response is that there are characteristics of those who refuse that may be different from those who accept and these may be different again from those that did not answer the phone; you simply do not know. By not having these people in your final sample, it may exclude people with their particular characteristics, leading to biased data. This problem is clearly one of missing data which is treated in detail in Chapter 14.

In dentistry we might hypothesize that when members of the general population are invited to be examined for an oral health study, those who are more anxious about dental care are probably less likely to agree, leaving the final sample distorted, again resulting in biased estimates. Losing all the anxious people may result in a very inappropriate service and squandered resource if your data are used.

There are two major things you can do to try to counteract the bias that will result from non-response:

1. Get the highest response rate possible. Because you know all about the people who respond, the risks of bias are hugely reduced where there is a high response rate. Time and resource preferentially targeted at ensuring a good response are well spent. There is a simple principle that says you are better with 90 people from a sample of 100 rather than 500 from a sample of a thousand. Although with the 90 people you will have lower

power and greater random error (represented by wider confidence limits) at least you can estimate how big that error is. With half of your sample of 1000 missing you could have a real problem with non-random error leading to bias, but you will never know where this is, so your data will be unreliable. Lots of very bad decisions are made on biased data with consequences for people and whole health care systems. There are cases where low response rates in small local surveys in the UK have resulted in dental services being set up, only for these to be underutilized by the population as a whole because the findings were based on a biased sample where regular attenders were grossly overrepresented.

2. Find out about the non-responders. If you know some of the characteristics of the non-responders you at least know something about the scale of the problem, and can sometimes weight the data to correct for known biases. If there are data on ages, genders or (better still) some sort of social or educational measure you can make an effort to adjust your final data to reflect this by weighting it back to how it would be in the population. In a dental study it is even worth trying to get information from the refuser on one or two items of key dental data at the time of refusal. What you need depends on what you are studying; it could be whether they have a dentist, whether they have natural teeth or whether they have had pain. This can be used to adjust your data to be more representative of your target population.

Bear in mind that post hoc correction of the data as described above (under point 2) is very much a second best, and can get rather messy. Get a good response first. Missing data is an important area in its own right and is covered in detail in Chapter 14.

7.3.5 Stratified samples and correcting for stratification

Sometimes a sample needs to be drawn where there is deliberate over sampling of certain subgroups. This allows less well represented subgroups to be included in the final sample in sufficient numbers to allow more precise estimates or to give greater statistical power to allow hypotheses to be tested relating to that group (see Chapter 10 on power considerations with balanced and unbalanced samples). An example of this would be in studies of older people. If you take a random sample of the whole population over the age of 65, the sample will be dominated by younger people because (obviously) people die as they age. A random sample of 1000 older men will give the numbers shown in Table 7.1 (based on recent UK population data). Clearly, if we wish to compare older groups with younger groups the oldest group (85+) will only have 94 people and this may not give us the power or precision we want. Estimates of proportion or means relating to this group will have wide confidence limits and the ability to compare with other groups and detect significant differences will be compromised.

To solve this problem a *stratified sample* can be drawn. A random sample, of the same overall size (say 1000) is drawn but each category is randomly sampled

Table 7.1 Table showing the size of a proportional sample of older men in different age groups (A), the stratified sample sizes (B) and correction weights for reporting data from the full sample (A/B) (modelled from UK National Census, Office for National Statistics).

Age group	Number in population	% of population aged 65+	(A) Number in sample of 1000	(B) Number in stratified sample	(A/B) Correction weight
65–69	1307300	30.74	307	200	1.535
70–74	1111400	26.13	261	200	1.305
75–79	862200	20.27	203	200	1.015
80–84	572400	13.46	135	200	0.675
85+	399700	9.40	94	200	0.470
All	4253000	100	1000	1000	

separately, in this case there are 5 age-bands targeting 200 people in each (again see Table 7.1). So long as data are reported from within age bands the stratified sample will give much greater power and precision from the older groups. A problem arises if you need to go back and report for the whole sample because your sample (in this case everyone aged 65 and over) will be heavily biased by the oldest participants because they were deliberately over-sampled. If for example an estimate of the mean number of teeth for everyone was required then you would end up with a mean that was artificially low because the oldest people will almost certainly have fewer teeth, pulling down the overall mean. Correcting for this is relatively easy by calculating a weight for each category which will be used as a multiplication factor to adjust the different groups up or down to restore the right proportions to represent the whole population. See Table 7.1 which gives the weight you would need to return your data to a state representative of the whole population aged 65 years or more. The relevant weight is applied to each person in the sample to increase or decrease the contribution that their individual data makes to the whole.

This principle of weighting is the same as that described in the previous section where it is used to correct for known biases resulting from sampling errors, but stratifying and re-weighting can get complicated. The more stratification factors involved the more weighting is required. Sticking with the older people theme, if you had men and women you would probably need to stratify by age and sex because women live longer than men. The correction weight for each person then is the product of their two weights; one for age group and one for sex. It is possible to stratify as much as you want but the errors for the overall weight will multiply.

7.4 Design issues: longitudinal studies, cohort studies

Longitudinal epidemiological studies can be very powerful tools indeed, but are extraordinarily difficult to run and usually require a long-term vision. A cohort is

recruited and then observed over a period of time. In a clinical trial an experimental element is introduced, but in a longitudinal epidemiological study the exposures are usually environmental, social or medical but the interaction between the exposure and time allows a much clearer indication of what may be a cause and effect relationship than is possible in a cross-sectional study.

The most difficult methodological problem with longitudinal and cohort studies is the maintenance of the sample throughout. For cohort studies the concept of a sampling frame and a representative sample are important in terms of the external validity of the findings. However the additional problem in longitudinal studies is sample attrition. Ten years into a very long-term study, it can be difficult to maintain contact with all of the participants; the effect is similar to poor response in a cross-sectional study in that you end up with non-random error and a final sample which may be biased.

Managing sample attrition in a cohort study is similar to managing non-response in a cross-sectional survey. The first priority is to retain as much of the sample as possible because if everyone is retained there will be no bias due to attrition. The second is to know about those who are lost. In a cohort study the latter is relatively easy as there will be a lot of information about the people who started in the sample, so it is possible to identify where any bias may be creeping in, and, if necessary, it can be corrected by weighting as described above.

The ultimate longitudinal studies are those which follow birth cohorts through a lifetime. There are a handful of such long-term samples around the world which provide a unique opportunity to identify the influences on health across a life-course. A recent example from New Zealand shows just how this can work and identifies caries trajectories that can be traced from an early age (Broadbent, Thomson & Poulson, 2008).

7.5 Design issues: case-control studies

Where you are interested in a condition that is rather uncommon you may want to sample just for people with that condition. People with the condition can then be compared with controls that do not have the condition and their characteristics and risk factors compared.

In this instance the sampling is relatively easy for the cases, provided that there is absolute clarity and consistency regarding the *case definition*. Even if the cases are a convenience sample (patients attending a clinic for example), by defining inclusion and exclusion criteria a priori, and ensuring that there is no further selection from those who fulfil the criteria then bias is eliminated, the sampling frame can be clearly defined and the external validity of the sample should be clear. A good response is also helpful where this is relevant.

Selecting controls is more difficult. The sample of cases will be dictated by the demography of the condition. People seeking treatment for Temporomandibular Disorder (TMD) for example will be several times more likely to be female than male. Figure 7.2 shows the distribution of cases and controls in a case-control

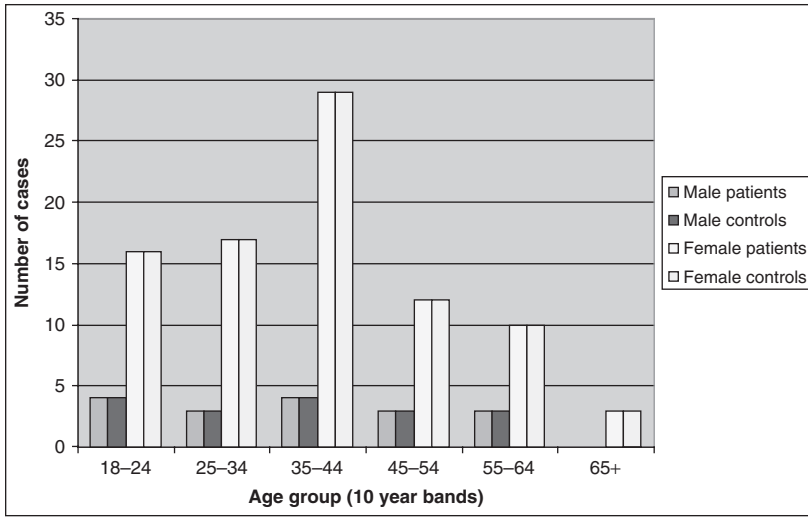


Figure 7.2 Plot showing matched cases and controls in a study of Temporomandibular Disorder and quality of life. Matching in this case was in 10-year bands and by both age and sex. Note the smaller number of males as the study was based on those seeking care. Because of the effects of both age and gender on reported oral health related quality of life, failure to match by both could have led to incorrect conclusions (courtesy of Dr A Moufti).

study of the impact of TMD on quality of life and illustrates this discrepancy. When selecting controls it is important to match these confounding variables to neutralize their effects. The normal strategy is to match for age and sex, but for certain conditions there may be other important confounders that might need to be matched. Clearly the more variables that are matched the more complicated it can become to find controls and care needs to be taken that cases and controls are not matched on a variable that is too closely related to the variable of interest (this is over-matching). Controls are therefore not random, but they are still drawn from a sampling frame of some description, often it is a convenience sampling frame, but it should be an appropriate one. Finally, it is crucial that any measurements made on the cases are made in exactly the same way as on the controls.

7.6 Bias in dental epidemiology

Bias is a fundamental concept in epidemiology. It can be defined as 'deviation of results or inferences from the truth, or processes leading to such deviation' (Grimes & Schulz, 2002). It is the result of errors occurring in the data that are not random but systematic and which can result in conclusions being drawn that are, as a result of the bias, invalid. There are a huge range of ways to classify bias and many different types and subtypes, most of which could find some application in

oral epidemiology. In the interests of simplicity we have used a classification here which includes only three comprehensive categories, all of which are relevant to the dental researcher.

Confounding bias Confounding bias has been discussed at the beginning of this chapter. It is where two or more variables occur together and where both (or all) may be risks for a disease. A dental example may be poor oral hygiene and smoking in the context of periodontal disease. Disease might be more common or more severe both in smokers and in those who have poor oral hygiene. However if smokers also tend to have poor oral hygiene then it is difficult to sort out whether the increased risk is to do with the smoking, the poor hygiene or a bit of both. Unless the confounding is managed through study design or analysis the estimate of the size of the risk for either of them may be biased, in this case inflated. There are ways this can be managed both by good design and analysis, but only if confounding variables are identified and included in the data collection. You cannot answer the question about smoking unless you have the data on the other risks!

Selection bias Selection bias is where the sample being used is, for whatever reason, not representative of the target population. This has also been discussed earlier in the chapter but can mean that the relationship between an exposure or risk factor and an outcome becomes distorted. This might be because of non-random sampling, or perhaps because the response rate is low ('response bias'), tending to overrepresent or underrepresent certain parts of the population. These issues are common but difficult or impossible to manage after the sample is achieved. Good sampling is fundamental.

Measurement bias This third form of bias is important in oral health research. Random errors occur all the time during data collection. In a sense these are not a major concern, though if they can be minimized (by careful measurement and good criteria) the precision of estimates and the ability to identify statistically significant differences between groups are enhanced. This is not measurement bias. Measurement bias occurs where measurement errors are systematic, not random, and again, the bias that results can easily result in flawed conclusions. A good dental example of this is a dental survey, cohort study, case-control study or trial where there are a number of dentists collecting clinical data, for example on dental caries. Different examiners do not always agree on what is a lesion and what is not. This is inevitable and every clinician can sympathize, but if the differences are systematic, and if one examiner persistently scores carious lesions where others do not then that individual will bias the estimates. If they are one of twenty examiners all seeing a random sample of participants, and the other 19 agree with each other, then there will be only a small overall error. If there are only two examiners then the overall error will be much higher but even that is not always a major problem if they are seeing the same types of participants. The really serious problem arises when the 'over-sensitive' examiner is restricted to one group of participants, for

example if they are the only examiner in one particular region or if they examine all of the people who have had one exposure (perhaps they are examining in the only fluoridated area). This can be disastrous as the inevitable conclusion is that disease is higher in that group than in any other. The policy consequences are easy to extrapolate. For this reason, training and calibration are a fundamental part of clinical epidemiology. The differences between examiners can be tested, if necessary retested, and then steps taken to minimize the risk (e.g. dropping an examiner, changing the allocation of subjects to examiners). See Chapter 12 for the evaluation of the performance of dental examiners vis-à-vis a gold standard and Chapter 16 for incorporating measurement error and misclassification into the statistical analysis.

One specific type of measurement bias of importance to oral health is recall bias. This is where retrospective questions are used, usually in a questionnaire. Different groups may recall their own experiences of similar events quite differently. For example nervous dental patients may recall experiences that others may completely forget, resulting in a conclusion that they have had a different experience to the rest of the population. Recall bias is difficult to manage but steps can be taken in terms of questionnaire design and validation that may help.

Overall, managing bias is arguably the most important issue in epidemiology, and bias can only really be managed by taking steps in advance. Once systematic errors are present in the data it can be almost impossible to get rid of them. Think ahead!

7.7 Summary

Epidemiological studies represent a powerful approach for identifying disease risks and possible solutions at a population level. Different approaches are required to answer different questions but in all cases, care with samples and (where appropriate) response or retention rates is critical.

References

Broadbent, J.M., Thomson, W.M., & Poulton, R. (2008). Trajectory patterns of dental caries experience in the permanent dentition to the fourth decade of life. *J Dent Res*; **87**: 69–72.

Dean, H.T., Arnold Jr, F.A., & Elvove, E. (1942) Domestic water and dental caries V. Additional studies of the relation of fluoride domestic waters to dental caries experience in 4425 white children aged 12–14 years, of 3 cities in 4 states. *Publ. Hlth. Rep* **57**: 1155–79.

Greenland, S. & Robins, J. (1994) Ecologic studies – biases, misconceptions and counterexamples. *Am J Epidemiol* **139**: 747–60.

Grimes, D. & Schulz, K. (2002) Bias and causal association in observational research. *Lancet* **359**: 248–52.

Kelly, M., Steele, J., Nuttall, N., *et al*. (2000) *Adult Dental Health Survey: Oral Health in the United Kingdom in 1998*. London: TSO.

Mason, J., Pearce, M.S., Walls, A.W.G., Parker, L., & Steele, J.G. (2006). How do factors at different stages of the lifecourse contribute to oral health related quality of life in middle age for men and women? *Journal of Dental Research* **85**: 257–61.

Pearce, M.S., Steele, J.G., Mason, J., Walls, A.W.G., & Parker, L. (2004) Do circumstances in early life contribute to tooth retention in middle age? *Journal of Dental Research* **83**: 562–6.

Pearce, N. (2000) The ecological fallacy strikes back. *J Epidemiol Community Health* **54**: 326–7.

Pitts, N.B., Chestnutt, I.G., Evans, D., White, D., Chadwick, B., & Steele, J.G. (2006) The dentinal decay experience of children in the United Kingdom in 2003. *Br Dent J* **200**: 313–20.

Riley, J.C., Lennon, M.A., & Ellwood, R. (1999) The effect of water fluoridation and social inequalities on dental caries in 5-year-old children. *Int J Epidemiol* **28**: 300–5.

Scannapieco, F.A., Bush, R.B., & Paju, S. (2003) Associations between periodontal disease and risk for atherosclerosis, cardiovascular disease, and stroke. A systematic review. *Ann Periodontol* **8**: 38–53.

8

Qualitative research

Christophe Bedos, Pierre Pluye, Christine Loignon and Alissa Levine

8.1 Introduction

The scientific world has historically been marked by the dominance of natural sciences over social sciences to the point that mathematics is sometimes labeled the 'queen of the sciences'. The criteria of scientificity from the natural sciences, such as objectivity, quantification and reproducibility, have ruled research fields that focus on human cultures and social phenomena. This dominance was criticized by a founder of sociology, Weber (1978), who stressed the differences between the natural and the social worlds and recommended the adoption of comprehensive approaches, an early form of qualitative research.

The long domination of 'numbers' in the social sciences, and public health in particular, has been further challenged in recent decades. Some scholars have emphasized the limitations of 'quantitative' research: it is mainly based on a deductive approach that does not favor the emergence of new perspectives; in addition, it often provides de-contextualized data in which human behavior is oversimplified (Guba & Lincoln, 1994). These limitations may explain the rise of another type of inquiry: qualitative research.

Creswell and Plano Clark (2007, p. 249) defined qualitative research as a 'process of understanding based on a distinct methodological tradition of inquiry that explores a social or human problem. The researcher builds a complex, holistic picture, analyzes words, reports detailed views of informants, and conducts the study in a natural setting.' The objective of this chapter is to provide an introduction to qualitative research and explain how it could be used in dental public health.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

8.2 A global perspective on qualitative research

8.2.1 A historical view on qualitative research

We will start our history of qualitative research at the beginning of the twentieth century. Denzin and Lincoln (2000) refer to the first part of the twentieth century as the 'traditional period', a time when several western anthropologists became famous for their work in 'remote' lands. Margaret Mead, for instance, shared the life of indigenous tribes in Samoa for several years and, using fieldwork techniques such as participant observation and interviews, described various aspects of their culture.

Then, from World War II to the 1960s–70s, qualitative research entered what has been named the 'modernist phase' because early qualitative researchers tried to formalize their methodological approach and make it fit the principles of quantitative research validity. This search for rigor and credibility was exemplified in 1967 in a landmark book, *The Discovery of Grounded Theory*. The authors, Glaser and Strauss (1967), described the different steps of an inductive approach to building theories and provided a structured guide for sampling, collecting data, and systematic analysis.

The search for scientific recognition led to important crises in the 1970s and 1980s. Indeed, some social scientists contested the adoption of principles of validity drawn from quantitative research; instead, they defended their scientific traditions and their own criteria of scientificity. This conflict between different schools of thought, known as the 'paradigm war', has, however, declined in recent years, and qualitative research, which was mostly confined to the social sciences, has progressively entered the field of health research (Denzin & Lincoln, 2000).

Over the years, qualitative research has evolved to encompass various traditions and approaches: Tesch (1990), for instance, identified 21 'types' of qualitative research. As it is beyond the scope of this chapter to describe them, we will briefly outline five relevant approaches suggested by Creswell (2007): ethnography, narrative, grounded theory, phenomenology, and case study. As Table 8.1 shows, these approaches originate from various fields and have different goals. Narrative, for instance, aims at exploring the life experience of individuals and relies on the stories that are told by them. In our field, this approach could be used to explore, within a life course perspective, the trajectory of people in the dental care system.

8.2.2 The philosophical foundations of qualitative research

Like most scholars, we believe that qualitative research cannot be understood 'without attention to the underlying assumptions [. . .] that guide the use of [any] particular research method' (Willis, Jost & Nilakanta, 2007). In our perspective, the researchers' *worldview*, or research *paradigm*, is important because it has an impact on how they define the problem, choose methods, and interpret data. Guba and Lincoln (1994) identified several coexisting paradigms, among which we will

Table 8.1 Five common approaches in qualitative research (adapted from Creswell & Plano Clark, 2007).

Approach	Fields of origin	Aim
Ethnography	Anthropology and sociology	Describe the shared culture of a group of individuals
Phenomenology	Psychology and philosophy	Understand a phenomenon and how it is experienced by people
Narrative	Humanities and social sciences	Explore the life experiences (biography, life history) of an individual or a group of individuals
Grounded theory	Sociology	Generate a theory (of a social process, an action) 'grounded in data from the field'
Case study	Human and social sciences	Study an issue through an in-depth description of a case

Table 8.2 Three different worldviews in social sciences (adapted from: Guba & Lincoln, 1994; Willis *et al*., 2007).

	Postpositivism	Critical theory	Constructivism
Nature of reality (Ontology)	Reality exists but is difficult to apprehend	Reality exists but is difficult to apprehend	Reality is multiple and 'constructed' by people who observe it
Nature of knowledge (Epistemology)	• Knowledge is an approximation of reality • Emphasis on objectivity	• Knowledge is mediated by values and ideologies • Emphasis on subjectivity	• Knowledge is constructed by the researcher and the subject • Subjectivity is greatly valued
Nature of methods (Methodology)	Preferably quantitative methods	Both quantitative and qualitative methods are acceptable	Preferably qualitative methods

succinctly describe three common ones: postpositivism, critical theory and constructivism (Table 8.2).

Postpositivism, which is probably the dominant paradigm in health research, is an evolution of positivism, first described by Auguste Comte in the nineteenth century. Postpositivist researchers assert that reality exists but, as it is difficult to

grasp, knowledge is only an approximation of reality. They also put the emphasis on objectivity and favor quantitative methods. Consequently, their use of qualitative methods is limited and generally serves to develop questionnaires.

The tenants of critical theory, which derived from Marxism in the first half of the twentieth century, focus on power relationships in society and aim for emancipation (Willis, Jost & Nilakanta, 2007). In contrast to postpositivist researchers, they accept an 'ideological bias' and favor subjectivity. Even though they use both qualitative and quantitative methods, they give preference to research strategies that are participative and empower the people under study.

Finally, constructivism stipulates that realities are multiple and socially constructed, and that there is no universal truth. In this perspective, constructivist researchers consider that subjectivity is unavoidable and even necessary. They therefore favor the use of qualitative methods.

The question of the researchers' worldview is important because it sustains the rules of scientificity. For example, differing views regarding the importance and meaning of objectivity and subjectivity are a potential source of disagreement between constructivists and postpositivists and may impede their collaboration.

8.2.3 The purpose of qualitative research

8.2.3.1 Usefulness and strengths of qualitative research

Researchers have recommended that qualitative research be used in dental public health (Hendricson, 2003; Bower & Scambler, 2007) because it is a powerful means of understanding complex social phenomena and people's perspectives on these phenomena. Indeed, one important difference with quantitative research is in the way the research questions are formulated: instead of asking 'how many' or 'how often', the qualitative researcher rather asks 'why' or 'how' in order to uncover people's perceptions (Bower & Scambler, 2007).

Qualitative research has important strengths. First, it produces rich data that standardized questionnaires cannot get at in as much depth. The researchers can immerse themselves in a culture and have direct interaction with the people under study, which enables them to apprehend people's behaviors and understand their perspectives on specific issues. Second, it allows the researchers to capture events in their natural setting and take the context into account, whereas quantitative approaches often decontextualize human behavior. It thus offers a holistic view of the phenomenon under study and helps identify issues that would have been ignored with a quantitative approach.

Qualitative research is therefore particularly useful in exploring phenomena about which little is known. Through an inductive process, it can help in developing theories and generating hypotheses. It is also helpful when quantitative research is unable to provide a satisfactory explanation for a phenomenon: qualitative research may then provide data that complement and enhance the results of quantitative

research. As Patton (2002, p. 10) states, qualitative research serves to 'illuminate the people behind the numbers and put faces on the statistics'.

8.2.3.2 An example of qualitative research

To illustrate the purpose and strengths of qualitative research, let us examine the issue of access to dental services among the poor. In Canada, as in other industrialized countries, quantitative studies have shown that the poorer the people, the lower the proportion that consult a dentist in a preventive manner (Bedos *et al*., 2004a) and the longer the delay before consulting after a dental problem occurs (Bedos *et al*., 2004b). This applies to welfare recipients, who tend to underuse professional dental services despite benefiting from public dental insurance that integrally covers most basic treatment. These quantitative studies provide important information with respect to the outcome – low use of dental services – but offer little knowledge regarding welfare recipients' process of accessing services, as well as their motivation to consult and the barriers they encounter.

Two studies exploring access to dental care among welfare recipients were recently conducted in Montreal, Canada to investigate these questions. The first was based on open-ended interviews with 16 people (Bedos *et al*., 2003), whereas the second relied on 8 focus groups in which 57 welfare recipients participated (Bedos *et al*., 2005). These studies revealed how people interpreted oral symptoms, how they dealt with them, how and when they decided to look for a dentist, and finally how they interacted with dentists and made treatment decisions. For instance, they interpreted the absence of symptoms as an absence of dental illness and perceived little need to consult preventively. They knew, however, that pain indicated a pathological process and signaled a need for professional treatment. But they preferred to wait and suffer because they deeply feared procedures such as local anesthesia. In this process of adapting to pain, they generally had recourse to analgesics combined with popular methods such as applying aspirin tablets to the surface of the painful tooth, massaging the jaw or drinking alcohol. These studies also showed barriers that welfare recipients encountered during the care episode. For instance, as endodontic treatments were not covered by public dental insurance, they often faced extraction of the painful tooth as their only solution. This would teach them to accept extraction as a reasonable option and even lead them to raise doubts about the efficacy of endodontic treatment.

In summary, these two studies provided a rich description of a complex oral health behavior among welfare recipients and detailed information concerning the beliefs, experience and fears that determined poor use of dental services. This information could hardly have been obtained through traditional quantitative research: self-completed questionnaires, even including open-ended questions, would not have elicited emerging issues such as welfare recipients' acceptance of tooth extraction and their progressive apprenticeship of rejection of endodontic treatments. These studies also show that quantitative and qualitative methods can be highly complementary.

8.3 Conducting qualitative research

8.3.1 Mixing qualitative and quantitative methods

Even though qualitative methods can be used independently, it may be pertinent to link them with quantitative methods in a single study or in the same program. Combining methods brings advantages to the researcher, who may benefit from both: qualitative methods provide an in-depth understanding of social phenomena whereas quantitative methods produce data that may be generalized to a population. In Figure 8.1, we describe four basic *designs* mixing both methods that are adapted from Steckler *et al*. (1992).

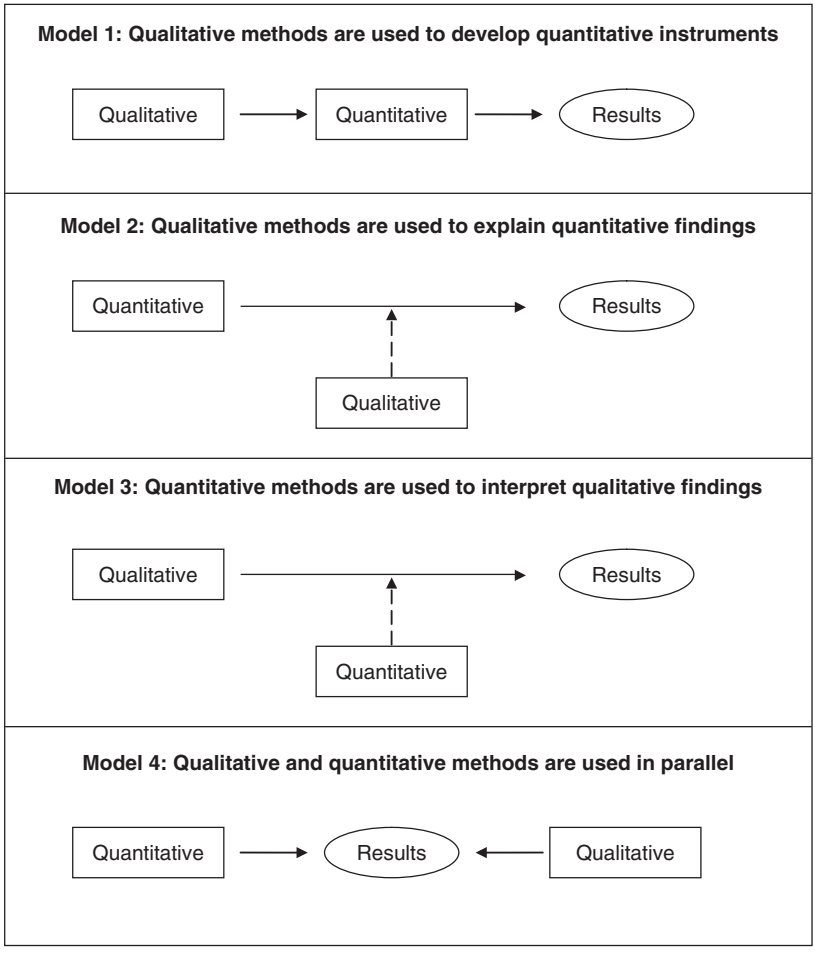


Figure 8.1 Designs mixing qualitative and quantitative methods (adapted from Steckler et al., 1992).

In Model 1, qualitative methods are used in an initial phase to develop quantitative instruments, such as questionnaires, that will then be the main source of data. Model 2 also includes 2 phases: the first phase, relying on quantitative methods, is followed by a qualitative phase whose purpose is to enhance the interpretation and meaningfulness of the initial results. This design is particularly useful when the quantitative phase generates unexpected results that the qualitative methods may help to explain. Conversely, in Model 3, quantitative methods are used to complement the results of the qualitative phase. These may take the form of a survey, for instance, which would help to generalize the qualitative findings. Model 4 is different from the previous three in the sense that the two methods are used concurrently: researchers may decide to place equal importance on both methods or prioritize one of them (Creswell & Plano Clark, 2007). The purpose of this design is to cross-validate the results: researchers 'analyze the results of each method separately and then decide if the results from each method suggest the same conclusion' (Steckler *et al*., 1992).

8.3.2 Sampling in qualitative research

8.3.2.1 Sampling strategies

Patton (2002) defined sampling in qualitative research as '*purposeful*' (or purposive), which differs from representative or random sampling in quantitative research: purposeful sampling means that the researcher strategically and purposefully selects cases that are information-rich. Under this broad term of purposeful sampling, Patton outlined 16 different strategies such as maximum variation sampling, homogeneous sampling, extreme case sampling and typical case sampling.

Each sampling strategy serves a particular purpose. Maximum variation sampling, for instance, aims at maximizing the diversity related to the research question, thereby describing the 'central themes that cut across a great deal of variation' (Patton, 2002, p. 235). For example, a researcher aiming at evaluating the implementation of a dental education program in schools will select a sample of schools that greatly differ in regard to outcome or characteristics, such as size or socioeconomic status. This could reveal how the program varies from one setting to another while at the same time showing what these programs have in common. Homogeneous sampling, on the contrary, aims at selecting cases – schools in our previous example – that are similar or share important characteristics. Extreme case sampling is another type of strategy targeting cases that are unusual and, for this reason, may provide rich and pertinent information. In our example, the researcher may select schools in which the dental education program is highly successful in order to identify factors of success. In contrast, typical case sampling focuses on what is average.

8.3.2.2 Sample size

Although there are no rules for sample size in qualitative research, most researchers rely on the principle of saturation. Saturation refers to the point at which additional

data do not improve understanding of the phenomenon under study: it is reached, for instance when new participants merely reiterate what has been said in previous interviews without contributing any further insights (Morse, 1995). This principle makes it difficult for the researchers to define the sample size before the start of data collection, which may create problems when writing the research proposal and planning the study's budget.

Qualitative researchers must therefore estimate the point of saturation based on their experience of the field and on the literature. However, authors who provide guidelines for sample size tend to disagree. For instance, Kuzel (1992) recommended 12–20 interviews for a maximum variation sample and only 6–8 for a homogenous sample, whereas Guest, Bunce and Johnson (2006) empirically evaluated 12 people as the saturation point for a homogeneous sample. Overall, most authors suggest a sample size of less than 50 people, which remains small compared to quantitative research (Guest, Bunce & Johnson, 2006).

When data is collected through focus group interviews, which rely on homogenous samples, the point of saturation may be reached quickly; a typical sample size is 3–5 groups (Morgan, 1998). Krueger (1998a, p. 72) explains that 'a rule of thumb has been to conduct three to four focus groups for a particular audience and then decide if additional groups (or cases) should be added to the study' according to the diversity of people's perspectives and the complexity of data.

8.3.3 Collecting qualitative data

Qualitative data generally come from three different sources, observation, open-ended interviews and documents, but we will only describe the first two because they are the most used in our field (Table 8.3).

8.3.3.1 Observing

The purpose of *observation* is 'to describe the setting [...], the activities that took place in that setting, the people who participated in those activities, and the meanings of what was observed from the perspectives of those observed' (Patton, 2002, p. 262). The degree of participation of the observer in the research setting can vary between two extremes: at one end, the observer acts as a participant in order to share in the study population's experience. This approach, called participant observation, has been widely used by anthropologists who aim to uncover different aspects of people's culture. At the other end of the spectrum, the non-participant observer remains a spectator and has limited interaction with the studied population.

Whether participant or not, observers produce 'field notes' that become the basis for analysis. While observing, they generally describe on a note pad the important aspects of the phenomenon under study, such as the setting and the actions that occur in this context. Right after the observation session, researchers complete and organize their notes, since it is difficult to write everything down as it is happening. This process may be enhanced by the use of pictures, audio or videotape records: pictures, for instance, facilitate the description of the research setting.

Table 8.3 Characteristics of the main data collection methods in qualitative research.

	Observations	**Semi-structured interviews**	**Focus groups**
Characteristics	Observation of the phenomenon in its natural setting: • Participant observation: the observer acts as a participant and experiences the phenomenon • Non-participant observation: the observer is a spectator	• Individual interview in which the researcher invites the participant to talk in depth on the subject. • The researcher uses an interview guide that includes the questions or themes to explore.	• Group interview (6–10 participants) 'focusing' on a specific subject • Conducted by 2 people: a moderator uses an interview guide and an assistant takes notes
Data collected and analyzed	Field notes (may be complemented by audio-visual material)	Transcript (verbatim) of audio-taped discussions	Transcript (verbatim) of audio-taped discussions
Strengths	Direct observation of the phenom enon in order to: • Identify factors that participants might not perceive as relevant • Collect data that people may be unwilling to discuss • Place the phenomenon in its natural context	• Capture what may be difficult to observe (opinions, emotions, beliefs, but also behaviors) • Help understand complex issues and obtain in-depth information	• Capture what may be difficult to observe (opinions, emotions, beliefs, but also behaviors) • Synergy among the participants generates data • The number of participants permits cross-validation of data

Table 8.3 (*continued*)

	Observations	Semi-structured interviews	Focus groups
Weaknesses and limitations	• Potentially time consuming • Technically challenging • The observer might be perceived as intrusive	• Technically challenging as the interviewer is part of the instrument • Provide 'indirect' information on the studied phenomenon	• Technically challenging and thus require a skilled moderator • Provide 'indirect' information about the studied phenomenon • Do not give much in-depth information about each participant

This table includes material presented by Patton (2002), Creswell (2007), Bower & Scambler (2007), and Morgan (1998).

This approach has several strengths. Indeed, instead of asking study participants to describe a phenomenon, the researchers directly observe it in its natural setting. This allows them to uncover routines that people might not pay attention to or might be reluctant to talk about in an interview. Another strength is that the researchers observe the phenomenon in its natural setting, helping them to understand the context and thereby providing a holistic perspective on human behavior. Observation also has disadvantages, in particular the fact it is time consuming and technically difficult for the participant observer (Fetterman, 1998).

8.3.3.2 Interviewing

Individual open-ended interviews A key element in interviewing is the degree of 'structure' of the interaction between the interviewer and the interviewee, which may vary from structured to unstructured. In structured open-ended interviews, the researcher reads a questionnaire in a standardized way but, as there are no predetermined response categories, the participant responds in his or her own words. On the opposite end of the continuum – unstructured interviews – the interviewer does not use any predetermined set of questions, which instead emerge from the conversation; this approach is generally used by ethnographers who know little about the phenomenon under study and adopt an inductive approach.

Probably the most common type of qualitative data collection method in health research, the semi-structured *interview* is situated between these two extremes. The researcher uses an interview guide that includes a list of questions but, unlike in

structured interviews, follow-up questions are developed and pursued during the course of the discussion to obtain deeper information. This approach allows the researcher the flexibility to explore unanticipated but relevant topics as they emerge from the discussion. The researcher is therefore often considered to be an integral part of the research instrument since the quality of the data relies greatly on his or her interviewing skills.

It is important to conduct a semi-structured interview in a quiet place in which the participant feels comfortable and is able to talk openly, such as in his or her own home. Usually, an interview lasts less than two hours but several additional sessions may be conducted if necessary. As it is difficult to take detailed notes during an interview, the discussion is generally audiotape recorded in order to be transcribed verbatim and then analyzed.

The main strength of semi-structured interviewing is that it allows the researcher to address what may be difficult to observe, such as perceptions, emotions, beliefs, and even behaviors. In addition, its flexibility allows for the freedom to explore complex issues and identify emerging themes or hypotheses.

Focus group interviews *Focus groups* are 1–2 hour-long group interviews that 'focus' on a specific subject of interest. In the field of health research, they have often been used to assess people's experience of illness, attitudes to health and health behaviors (Kitzinger, 1995).

A focus group is generally composed of 6–10 participants having, in principle, a shared experience of the studied phenomenon. It is directed by a moderator and an assistant moderator. The moderator uses an interview guide that usually includes 2–5 'key questions' (Krueger, 1998b), and encourages all participants to interact and exchange views instead of responding in turn. In parallel, the assistant moderator writes detailed notes during the discussion and assists with the logistics. Unlike individual interviews, which are often held in informal settings, focus groups require adapted facilities, preferably a comfortable room in which people sit around a table.

One of the strengths of focus groups is related to the dynamics among participants: building on each others' comments, participants are often able to clarify their views in a way that would be difficult to foster in individual interviews. Each focus group session also provides the perspectives of several people at the same time, which improves the quality of data by providing a 'check and balance' of the different viewpoints (Krueger & Casey, 2000).

Although the number of participants provides several advantages, it also brings limitations as each participant may be able to speak merely a few minutes in all; as a consequence, focus groups do not provide much in-depth information regarding each participant's experiences. Another weakness is that focus groups require a skilled moderator able to manage a group discussion; for instance, the moderator must ensure that talkative participants do not take over and that shy ones are drawn into the discussion.

8.3.4 Analyzing qualitative data

The way researchers analyze qualitative data greatly varies according to their worldview and to the approach they choose. In this section, we will present a generic method that is inspired by Miles and Huberman (1994). It includes three main phases: data organization, data display, and data interpretation. It is important to mention that, in contrast to quantitative data analyses, qualitative analyses are quasi simultaneous to data collection: each interview is quickly transcribed and analyzed in order to prepare the next interview and even guide the selection of subsequent participants. Furthermore, the three phases of data analysis – data organization, display, and interpretation – are concomitant and influence each other in a recursive process.

8.3.4.1 Data organization: the coding process

Coding is the process of categorizing qualitative data. It consists of cutting a text into meaningful segments and assigning codes to these segments: each code represents an idea or a concept, and it labels all parts of the text that refer to this idea such as sentences or paragraphs. For instance, in a study focusing on the quality of dental care, a code entitled 'fear of dental procedures' would be used to identify each passage of the transcribed interviews relating to this issue.

The coding process can be conducted inductively. With the grounded theory approach, for instance, researchers let the codes 'emerge' from the text, which means that they identify them while they read the transcribed data. The coding process can also be more deductive: some researchers start with a list of codes that is determined a priori in relation to the research questions. However, it is generally necessary to refine or reconceptualize the codes during this process: some are split into subcategories, others are mixed. Coding is thus an iterative process that, especially in its early stages, obliges the researcher to review data that have been previously coded in order to systematically apply newly developed codes.

Codes allow researchers to efficiently consult the data and retrieve specific segments of the text according to their meaning. For instance, the code 'fear of dental procedures' that we cited earlier would permit us to retrieve and display all pieces of all transcribed interviews on that subject. Because qualitative research generates a large amount of data, coding is an extremely demanding process: consider, for example, that each one-hour interview may provide about 30 pages of text; this means that a sample of 20 people typically yields 600 pages to be coded.

8.3.4.2 Data display

Data display means presenting coded data in a condensed format in order to enhance interpretation and draw conclusions. The most common form of display is textual: the researcher presents the segments of text under each code or summarizes them. For instance, the retrieval of the code 'fear of dental procedures' may result in 25 pages of transcribed text for the whole sample; the researcher, however, may

prefer to reduce and summarize it for each person interviewed or for the sample in general.

The other forms of data display are more visual and include matrices, graphs, and networks. Matrices, for instance, are tables in which each column may correspond to a code and each row to a case – a site under observation, a person interviewed or a focus group. By filling the cells with brief notes, the researcher provides a compact summary of the data collected that facilitates comparison among the cases. For instance, in the column corresponding to the code entitled 'fear of dental procedures', the researcher may briefly describe, row by row, the perspective of each person interviewed on this subject.

These visual forms are highly recommended by Miles and Huberman (1994, p. 92), according to whom 'the chances of drawing and verifying valid conclusions are much greater than for extended text, because the display is arranged coherently to permit careful comparison, detection of differences, noting of patterns, seeing trends and so on'.

8.3.4.3 Data interpretation

The purpose of this phase is to interpret the data that are displayed in texts or matrices. As we mentioned earlier, however, qualitative research relies on flexibility, which means that sampling and data collection are concurrent with analysis; interpretation occurs as soon as data collection starts, but the initial interpretations are progressively refined and the conclusions gain in precision as the study advances.

Miles and Huberman (1994) described 13 tactics for interpreting data such as noting patterns, clustering data, or identifying relations between codes. One of the most common, noting patterns, consists of detecting regularities among participants. In carefully examining a matrix, for example, the researcher may notice that all people interviewed greatly fear injections at the dental office; hence, fear of injections would be a pattern, or shared issue among the participants. Clustering data is another tactic that resides in grouping codes or participants according to meaningful categories. For instance, participants may be classified into 'fearful' and 'non-fearful' categories with respect to dental procedures. It is also possible to identify associations between two or several codes, such as 'past experiences at the dental office' and 'fear of dental procedures'.

8.3.4.4 Computer assisted qualitative data analysis software (CAQDAS)

Even though researchers traditionally analyzed qualitative data 'by hand', most now utilize software programs. Yet unlike statistical programs such as SAS or SPSS, qualitative software does not perform the analysis but rather assists the researcher in coding and retrieving the data. The researcher using CAQDAS thus maintains control of the analytical process.

Choosing software is not easy because many have been developed in recent years and there is no consensus on the 'best CAQDAS'. As a consequence, researchers should compare the different products and determine the most appropriate according to criteria such as their analytic approach, the type of data collected,

or the cost of the package (between $0 and $1000). A comprehensive and regularly updated review is provided by the 'CAQDAS Networking Project' and is available at http://caqdas.soc.surrey.ac.uk.

8.3.5 Validity of qualitative research

The criteria to assess research also vary according to the researchers' worldview and to the approach they choose. Internal and external validity, which are used for quantitative research within the postpositivist paradigm, must be significantly adapted if they are to apply to qualitative research. Lincoln and Guba (1985) suggested *credibility* and *transferability* as alternative criteria.

For qualitative researchers, credibility is a crucial issue with respect to validity: research is credible when the results are plausible for those who participated in the study and those who read the research results. Credibility relies on rigor and can be enhanced by a series of procedures that are cited in Table 8.4. Prolonged engagement, for instance, signifies that the researchers – the observers in particular – are involved in the field for a sufficient amount of time to build trust

Table 8.4 Procedures to enhance credibility and transferability in qualitative research.

Credibility (alternative to internal validity)	**Transferability** (alternative to external validity)

• Prolonged engagement of the researcher in the field
• Control or take into account the effects of the interaction between the researcher and the participants:
  • Reflexive journal completed by the researcher (notes that critically describe researcher's own thoughts, decisions, or biases over the course of the project)
  • Peer debriefing with co-researchers
• Triangulation (comparison of results obtained by different data sources, methods of data collection, and researchers).
• Checking the interpretations with the participants after data collection
• Rigorous coding of data
• Analysis of 'negative cases' or outliers (cases that do not fit with the main results)
	• Thick description of the sampling strategy, the context, the methods, and the results, thereby allowing others to extrapolate findings or to reproduce the study in a similar or different setting.

Table 8.5 Overall comparison between qualitative and quantitative research.

	Quantitative research	Qualitative research
Purpose	Explain, compare and generalize	Understand, explore, generate hypotheses
Research questions	How much? How many? How often?	How? Why? What?
Sampling	Representative sampling (probabilistic)	Purposeful sampling (select 'information-rich' cases)
Sample size	Large	Small
Data collection	• Use of validated instrument (such as structured questionnaires) • Controlled setting is ensured or assumed • Minimizes the impact of the observer on the outcome (objectivity)	• The researcher is an integral part of the instrument and uses various techniques (such as interviews and observation) • Control of the environment is neither possible nor desirable • Generally values the interaction between the researcher and the participants (subjectivity)
Data	Numbers	Words
Analysis	• Occurs after data collection • Simplification of results • Imposed, predetermined categories	• Occurs throughout all stages of research • Examination of complexities to deepen comprehension of a phenomenon • Categories may emerge from the data
Strengths	• Indicates the prevalence of a phenomenon • May serve to predict future phenomena • Data collection and analysis can be conducted on a large scale in a limited amount of time	• Produces rich, detailed data • Provides a holistic view of the phenomenon under study and takes the context into account

Table 8.5 (*continued*)

	Quantitative research	Qualitative research
Weaknesses	• The results may be context-bound. For instance, social research in a controlled setting may not reflect actual behavior • It is difficult or impossible to account for unusual or conflicting results	• Limited ability to predict and generalize • Data collection and analysis is technically difficult, time consuming and therefore potentially expensive

with the participants and gain a deep understanding of the phenomenon under study. Other procedures, such as conducting peer debriefings with co-researchers during data collection, allow the researchers to better control the effects of the interaction with the participants and limit potential biases; such sessions force the researchers to analyze their possible influence on the participants while helping them to identify and control their own reactions and biases during fieldwork. This being said, it is important to remember that the definition of bias depends on the researcher's worldview: constructivist researchers, for instance, consider bias to be inherent to their approach whereas postpositivists try to minimize the influence of the interviewer on the participant.

Qualitative research cannot be judged according to standard norms of external validity, or generalizability, because it relies on samples that are small and not representative of the population. Instead, qualitative results may be 'transferred' to other contexts where others may estimate how the results of a study pertain to a different setting. In other words, the readers assess the similarities and the differences between two environments and evaluate the degree of transferability of the results, or what is transferable and what is not. Because this process is performed by readers, it is only possible if the research reports include a 'thick description', that is, a highly detailed account of the methods and the context in order to allow adequate comparison.

8.4 Conclusion

In this chapter, we have shown that qualitative research differs from quantitative research in many ways, as Table 8.5 summarizes. These differences may be considered a problem, which was the case during the paradigm war, but they may also be seen as a strength. For our part, we believe that quantitative and qualitative approaches are not only compatible, but complementary, as evidenced in mixed methods studies. We also believe that qualitative research should be more widely used in public health dentistry because it provides unique insights about people's behaviors, perceptions and beliefs.

It is important to emphasize that conducting qualitative research is neither easy nor quick: it requires time and effort to collect and analyze the large amount of data; it also demands skilled and experienced researchers to perform or supervise these tasks. Novice qualitative researchers should not underestimate the challenge posed by this complex but valuable approach. We invite them to deepen their knowledge with additional readings. We also recommend that they find mentors who will guide them at every step of the process.

References

Bedos, C. *et al*. (2003) The dental care pathway of welfare recipients in Quebec. *Soc Sci Med* **57**(11): 2089–99.

Bedos, C. *et al*. (2004a) Social inequalities in the demand for dental care. *Rev Epidemiol Sante Publique* **52**(3): 261–70.

Bedos, C. *et al*. (2004b) Dental care pathway of Quebecers after a broken filling. *Community Dent Health* **21**(4): 277–84.

Bedos, C. *et al*. (2005) Perception of dental illness among persons receiving public assistance in Montreal. *Am J Public Health* **95**(8): 1340–4.

Bower, E. & S. Scambler (2007) The contributions of qualitative research towards dental public health practice. *Community Dent Oral Epidemiol* **35**(3): 161–9.

Creswell, J.W. (2007) *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. 2nd edn. Thousand Oaks: Sage Publications.

Creswell, J.W. & V.L. Plano Clark (2007) *Designing and Conducting Mixed Methods Research*. Thousand Oaks, Calif.: Sage Publications.

Denzin, N.K. & Y.S. Lincoln (2000) The discipline and practice of qualitative research. In: N.K. Denzin & Y.S. Lincoln (eds), *Handbook of Qualitative Research*, Sage Publications: Thousand Oaks, Calif., pp. 1–28.

Fetterman, D.M. (1998) *Ethnography: Step by Step*. 2nd edn. *Applied Social Research Methods series*, vol. v. 17. Thousand Oaks, Calif.: Sage Publications. x, 165.

Glaser, B.G. & A.L. Strauss (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago,: Aldine.

Guba, E.G. & Y.S. Lincoln (1994) Competing paradigms in qualitative research. In: N.K. Denzin & Y.S. Lincoln (eds), *Handbook of Qualitative Research*, Sage Publications: Thousand Oaks, Calif., pp. 105–16.

Guest, G., A. Bunce, & L. Johnson (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods* **18**(1): 59–82.

Hendricson, B. (2003) It all starts with questions. *J Dent Educ* **67**(9): 965–9.

Kitzinger, J. (1995) Qualitative research. Introducing focus groups. *BMJ* **311**(7000): 299–302.

Krueger, R.A. (1998a) Analyzing and reporting focus group results. *Focus Group Kit Vol 6*, ed. D.L. Morgan *et al*. Thousand Oaks, Calif.: SAGE Publications.

Krueger, R.A. (1998b) Developing questions for focus groups. *Focus Group Kit Vol 3*. ed. D.L. Morgan *et al*. Thousand Oaks, Calif.: SAGE Publications.

Krueger, R.A. & M.A. Casey (2000) *Focus Groups: A Practical Guide for Applied Research*. 3rd edn. Thousand Oaks, Calif.: SAGE Publications.

Kuzel, A. (1992) Sampling in qualitative inquiry. In: B. Crabtree and W. Miller (eds), *Doing Qualitative Research*, Sage: Newbury Park, CA, pp. 31–44.

Lincoln, Y.S. & E.G. Guba (1985) *Naturalistic Inquiry*. Beverly Hills, Calif.: Sage Publications.

Miles, M.B. & A.M. Huberman (1994) *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks: Sage.

Morgan, D.L. (1998) Planning focus groups. *Focus Group Kit Vol 6*, ed. D.L. Morgan *et al*. Thousand Oaks, Calif.: Sage Publications.

Morse, J.M. (1995) The significance of saturation. *Qualitative Health Research* **5**(2): 147–9.

Patton, M.Q. (2002) *Qualitative Research and Evaluation Methods*, 3rd edn. Thousand Oaks, Calif.: Sage Publications.

Steckler, A. *et al*. (1992) Toward integrating qualitative and quantitative methods: an introduction. *Health Educ Q* **19**(1): 1–8.

Tesch, R. (1990) *Qualitative Research: Analysis Types and Software Tools*. Bristol, PA: Falmer Press.

Weber, M. (1978) *Economy and Society: An Outline of Interpretive Sociology*. Berkeley and Los Angeles, California: University of California Press.

Willis, J.W., M. Jost, & R. Nilakanta (2007). *Foundations of Qualitative Research*. Thousand Oaks, Calif.: SAGE Publications.

9

Data validity and quality

Finbarr Allen and Jimmy Steele

9.1 What do we mean by validity and quality?

Collecting data does not, in itself, make for good research. To answer a research question requires data that are clean, accurate and valid. Valid data describe accurately what we want them to describe or what we think they are describing.

Once the complexities of sampling, ethics and expense have been addressed, clinical data such as the number of teeth or fillings in a sample of the population are relatively easy to collect. Their validity depends on correct sampling, accurate measurement according to tight and appropriate criteria, good agreement between different observers in the application of these criteria, and, finally, the way in which data are subsequently summarized and presented. If all of these issues are addressed early and correctly, the clinical data should, hopefully, be valid. These issues will be discussed in a little more detail in the next section.

There are other validity issues to consider. In the rush to write a funding application or start data collection, researchers sometimes forget to step back and ask the really important question about data validity: Will my (technically valid) data answer the research question I am asking? Sometimes even that research question needs to be revisited before you can move on. For example, collecting clinical data of the highest technical quality allows us to describe the population in terms of pathological process, but not necessarily how oral health affects people on a day to day basis. The argument for trying to measure oral health and the effects of interventions using variables that capture the impact of oral health on people's lives rather than (or in addition to) just measuring clinical disease, is compelling. The measurement tools that allow you to do this are very different from the familiar clinical measures of DMFT or number of teeth, and the concerns regarding their

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

validity are quite different. This philosophical issue will form the later sections of this chapter.

First though there are some routine considerations relating to validity and quality of any oral health data that merit consideration.

9.2 The very basics of data quality

There is always a process by which your data finds its way from the research subject to the database. This will vary from project to project, but the consequences of ignoring this are potentially catastrophic. Methodological aspects of data collection and storage are considered elsewhere, and the following paragraphs should be read in conjunction with Chapters 5 and 7.

9.2.1 Clinical data

9.2.1.1 Clinical criteria

In order to measure any clinical condition the researcher needs robust, tried, tested and preferably widely used criteria. Fortunately the major clinical conditions such as caries, periodontal disease, tooth wear and even temporomandibular disorders (TMD) have been measured many times before, so there are well-established criteria which can be applied to any population. Sometimes there is a choice. In that case, look for those with a strong methodological and scientific background, examples may be the ICDAS (Ismail *et al*., 2007) system for caries or the Research Diagnostic Criteria for TMD (Dworkin & LeResche, 1992). Both of these systems have been developed from a strong theoretical base and have been subject to extensive research and testing. Criteria such as these sometimes need to be adapted to the needs of an individual study, and that is perfectly acceptable provided that great care is exercised to retain the theoretical structure. Rarely, a researcher needs to develop brand new clinical criteria. But if that is a road that you need to go down then take it seriously, base them on strong theoretical understanding and ensure proper piloting and preferably validity checking. Usually though the literature should help find an appropriate set of criteria.

9.2.1.2 The safe use of criteria

Any criteria are only a starting point, which still need to be interpreted and applied by the person collecting the data. The greatest problem for the researcher is getting different people (usually dentists or dental hygienists) to use the criteria in the same way. Where there is only one person collecting the data one does not need to worry about variations between examiners, but there are still two major risks.

1. The person applying the criteria may not apply them appropriately. This may be an issue of understanding the criteria, but more often relates to interpretation. For instance, a noncavitated discoloured tooth surface may

or may not have a carious lesion present and doubt may exist as to how this should be recorded. Failure to agree a consistent approach to an issue such as this will impact significantly on data quality. It is wise to calibrate the examiner with an experienced 'gold standard' prior to data collection to make sure that there is some control over how they are applied.

2. The person applying the criteria wavers in the way they are applied over the period of data collection. This is less easy to address but the normal approach is to ensure that a small number of cases are re-examined at two different time points to allow a measure of calibration to be applied to see to what extent the two examinations agree (see below).

Where more than one examiner is involved in data collection these two problems still apply, but there is an even greater problem. Namely, how do you get different examiners to interpret and apply criteria in the same way? In a clinical trial, where the differences in outcomes between two interventions or products may be small, very high levels of calibration are demanded to maximize the power to find a difference between treatments.

In a large clinical study such as a population survey or a case-control study, the issues are a little different. Indeed, the range of data being recorded is usually much larger, as is the sample size. It is still important to try to obtain the best agreement possible. For example one rogue examiner who examines only in a single discrete geographical area and who systematically scores caries much higher or lower than other examiners could result in completely spurious conclusions being drawn about the distribution of disease and entire policy decisions being made inappropriately. In other words, there is a risk of bias resulting from poor calibration. See Lesaffre, Mwalili and Declerck (2004) for an example whereby the impact of differential scoring of several examiners has been evaluated. Such problems can be minimized by good training and then identified and managed by calibration. Other steps can help to limit the impact of less than perfect calibration. For example the examiners could be mixed to ensure that they are subjected to a broad range of the population or of the number of examiners could be increased. Indeed, while multiple examiners are much more difficult to calibrate, the risks of one discrepant examiner possibly biasing the results are reduced, assuming the group of examiners are well trained. It is worth noting that one of the greatest limits to calibration when collecting clinical data from people is the tolerance of the volunteer patients to repeated examination; this is particularly problematic for periodontal data, and gives the whole process of training and calibration an ethical dimension.

Measuring calibration is difficult enough (see below), but the problem of managing poor calibration is just as challenging. If you find it you have to do something about it. There are two basic approaches. The usual approach is to provide further training and recalibrate, only dropping the examiner if he/she fails to reach an acceptable level of calibration. To do this the researcher has to have the time, space and infrastructure, and this is not always available. The alternative is to take the sometimes harsh step of dropping a discrepant examiner. This is not easy for many reasons, not the least because it leaves the study an examiner down, but the

problem has to be dealt with. The message really is that good training is paramount and will minimize the risk of poor calibration.

9.2.1.3 Measurement of calibration

The statistical approach to measuring agreement will depend on circumstances, but again professional advice should be sought as there are a number of scores which can be calculated to indicate levels of agreement.

Cohen's *kappa* κ is one of the most widely used measures. For a general reference on kappa statistics, the reader is referred to Altman (1999) for further details. In its simplest form it measures disagreement between two examiners for binary scores and is calculated from the observed and expected frequencies as $\kappa = (p_o - p_e)/(1 - p_e)$. The kappa score lies theoretically between -1 and 1, but in practice it is in between 0 and 1, indicating the extent to which different observers agree, with 1 being perfect agreement. The need for perfection will determine the minimum acceptable score but anything above 0.4 is sometimes considered acceptable, above 0.8 is excellent (Landis & Koch, 1977). It is probably wise to set limits for kappa 'a priori' then apply them ruthlessly. For ordinal scores a weighted kappa has been suggested, which accounts for the size of the discrepancy between scores. But also a kappa has been suggested for measuring the disagreement when giving nominal scores and a kappa evaluating scores between multiple examiners.

What a kappa score will not tell you is whether any disagreement is actually all because one examiner is consistently scoring very high or very low. This does not matter if you are aiming for a very high score anyway (agreement is agreement after all). However, in the process of identifying how to manage a discrepant examiner it may be useful to know how they are behaving, so the use of additional approaches may be indicated. This could be done by e.g. calculating sensitivity and specificity vis-à-vis a benchmark scorer – see also Chapter 12.

The process of *calibration* is always challenging and there is always some level of compromise, but the discipline of doing it is important as it forces a data quality agenda and it ensures that the researcher provides tight criteria and good training.

9.2.2 Getting data from the person to the database

This sounds very mechanical, but is critical. There are a number of ways of getting data onto a file, but generally the fewer the number of steps the less is the risk of error. The people who write down the codes, lead the questionnaires, and enter the data onto a computer are all important and need to be properly trained, whatever the nature of your data. It is also a good idea to pilot any data form for data entry before using it; the person entering the data will do it more quickly, cheaply and accurately if the form is designed well. At the data entry stage, double entry is often used for paper data to minimize the risk of errors as data are transferred to electronic form. Scores that do not agree between first and second entries are identified and can

be checked, radically reducing keystroke errors. There are too many possibilities to mention here, but cutting corners will result in mistakes and invalid data. This is dull but necessary. More recently, electronic data capturing methods have been used and the same basic principles of data collection and management apply. In theory, these systems can be programmed to monitor for double entry or incorrect entry, but the investigator still needs to be vigilant!

9.2.3 Missing data

Even the most carefully collected dataset is rarely complete; data are missing for a myriad of perfectly good reasons. The ability to distinguish between types of missing data is important. For example, where a tooth is missing there can be no data about its condition but you know it is missing. This is quite different from simply having no information at all, which in turn is quite different from a tooth which is present but for which it was impossible to determine its condition for some reason. It is important to be able to distinguish between these cases. The same principle can be applied to almost any type of data. When a subject is completing a questionnaire there may be good reasons why a single question may be left out. Perhaps it does not apply to the individual (for example a question about a denture only applies if you have one), perhaps they don't understand it or maybe they just forgot. Being able to distinguish these cases can be important because sometimes the type of missing data can be informative for the statistical analysis. Understanding the logic which underpins different types of missing data and coding them properly is a key step in preparing and coding the data file.

True missing data, where there is no indication why information is absent, is a real problem for the statistical analysis. Inappropriate dealing with missing data can severely bias the results of the study and missing data on covariates can seriously hamper the application of multivariate analyses (see Chapter 11). Every step should be made to avoid missing data, from double checking data forms before leaving the subject, to correct routing and coding for clinical and questionnaire data. However, missing data are almost inevitable in many areas of work. This is covered in much more detail in Chapter 14.

For specific types of data, for example a quality of life questionnaire, it may be important to combine scores for several variables (in this case questions) into a summary score. For example, the Oral Health Impact Profile (Slade & Spencer, 1994) comprises 49 questions, and a single missing code could mean that a total score cannot be collected for that subject. Because single missing codes are quite common one could theoretically end up with almost no cases to analyse (as with multivariate analyses). In these cases it is important to set some rules. Where only a few codes are missing it is often reasonable to 'impute' data, in other words to insert a code which represents a good 'guess' of the missing code. A popular approach is using the technique of multiple imputation, see Chapter 14. Where larger amounts of data are missing then the case may still have to be lost, but rules set a priori will allow with this problem to be dealt with consistently.

9.3 Patient self-reported data

9.3.1 Why should we include self-reported measures?

Objective measures of oral disease status such as the Decayed, Missing and Filled Teeth (DMFT) and the Basic Periodontal Examination (BPE) indices provide important data on tooth loss trends and markers of tooth loss. However, they are measures of stages in pathological processes such as dental caries, and provide no insight into psychosocial impacts of disease on daily living or *health-related quality of life*. It is clear that a greater insight into the consequences of oral disease can only be made using self-reported health assessment as well as collecting objective disease measurement data. Furthermore, as patients' assessment of their health-related quality of life is often markedly different to the opinion of health care professionals (Slevin *et al*., 1988), patient assessment of health care interventions is warranted. Uses of health-related quality of life measures have been described by Fitzpatrick *et al*. (1992), and are shown in Table 9.1

9.3.2 Choosing patient rated health status measures – how should this be approached?

As with clinical data collection, it is vitally important to have a systematic approach to choosing patient rated health status measures. If a researcher chooses an inappropriate measure, then conclusions of the research study may be spurious. The sophistication of measures currently available varies widely, and a number of theoretical issues need to be considered when selecting a health status measure. Broadly speaking, the main requirements of a health status measure have been summarized by Fitzpatrick *et al*. (1992). These are: (1) multidimensional construct; (2) reliability; (3) validity; (4) sensitivity to change; (5) appropriateness, and (6) practical utility. In essence, the measure chosen must be suited to the purpose for which it is being chosen, and have appropriate measurement properties. This issue has been addressed by Guyatt *et al*. (1993), who stated that some measures were suited to measuring between group differences in cross-sectional population studies, whereas others were more suited to measuring change in clinical trials and intervention studies. It must not be assumed that all health status measures can perform both tasks equally well. This principle is illustrated in Tables 9.2 and 9.3. In this study by Allen and Locker (2002), oral health related quality of life outcomes for

Table 9.1 Uses of measures of health-related quality of life

- Screening and monitoring for psychosocial problems in individual patient care
- Population surveys of perceived health problems
- Medical audit
- Outcome measures in health services or evaluation research
- Clinical trials
- Cost-utility analysis

Table 9.2 Comparison of mean baseline summary OHIP-EDENT, OHIP-14 and OHIP-49 scores, by group using unpaired t-test

	IG (n = 21)	CDG(n-33)	P
OHIP-49	101.67	55.80	<0.01
OHIP-14	30.20	15.1	<0.01
OHIP-EDENT	50.6	29.5	<0.01

Table 9.3 Changes in OHIP-EDENT, OHIP-14, OHIP-49 summary scores following treatment, by group using paired t-tests on mean diff = mean of pre-operative scores – post-operative score

Group	Postoperative mean (SD)	Mean diff (SD)	P	Effect size
IG (N = 21)				
OHIP-EDENT	38.7 (22.8)	11.8 (19.7)	0.02	0.9
OHIP-14	30.2 (10.3)	7.6 (13.7)	0.03	0.3
OHIP-49	71.19 (49.8)	30.5 (46.6)	<0.001	1.0
CDG (N = 33)				
OHIP-EDENT	23.1 (15.2)	6.4 (17.1)	0.04	0.4
OHIP-14	11.5 (9.4)	3.7 (13.9)	0.15	0.2
OHIP-49	40.6 (29.3)	14.8 (35.0)	0.02	0.5

edentulous patients treated with implant retained bridges ('IG') and conventional dentures ('CDG') were compared. The outcome measure used was the 49 item Oral Health Impact Profile (Slade & Spencer, 1994). For the purpose of analysis, the measurement properties of this measure were compared with two subsets of the OHIP, namely, the short version OHIP (Slade, 1997) and a subset of OHIP items (OHIP-EDENT) determined using an item impact method (Allen & Locker, 2002). The a priori hypothesis was the subjects who received implant retained prostheses would report a greater improvement than those who received conventional dentures. Table 9.2 compares baseline OHIP summary scores for each group using the three outcome measures.

These data indicate that, as expected, there were significant baseline differences between the groups and this was detected by all three outcome measures. Accordingly, this suggests that all three measures have good discriminant validity properties. The data in Table 9.3 illustrate within group pre/post treatment differences from the same study. The magnitude of pre/post treatment differences was determined using effect size statistics. Using benchmarks provided by Cohen (1977), an effect size of <0.2 is considered a small change, 0.2–0.7 a moderate change and >0.7 is a large change.

These data indicate that the OHIP-14 was not responsive to change in this study, and unlike the OHIP-49, was unable to detect clinically meaningful change.

This was attributed by the authors to floor effect (see Section 9.2.5 for explanation). The OHIP-EDENT had better responsiveness properties, and detected a clinically relevant change.

At a purely practical level, the fewer the questions in a health status measure, the greater its practical utility. So called 'single-item' or global measures have practical advantages, but fail to meet the requirement of multidimensional construct (e.g. contain a number of domains such as physical, psychological and social domains). In an oral health context, a single item measure such as 'In general, are you satisfied with the comfort of your mouth?' may be a broad indicator to satisfaction, but fails to yield insight into underlying causes of satisfaction/dissatisfaction. They are, therefore, unlikely to detect small or even moderate changes in clinical trials. Visual Analogue Scales (VAS) are also widely used as symptom scales, particularly in pain research. This involves a respondent rating their symptom experience by placing a vertical mark on a 100 mm line with extremes (e.g. 'no pain' and 'extreme pain') at either end of the line. The researcher then measures the distance along the line to the marking made by the patient. It is argued that use of a continuous scale affords a high degree of precision, but it seems unlikely that the underlying attribute can be measured to that level of precision. In addition, a VAS measures a single construct, and reliability of the measure is low (Streiner & Norman, 2003).

Locker (1988) raised awareness of the need for a conceptual framework for oral health outcome measures, and proposed a theoretical model for measuring oral health. This model attempts to capture possible consequences of disease, ranging from physical impairment (i.e. loss of an anatomical structure) to broader psychosocial consequences such as handicap. The model serves as a framework for development of multidimensional health status measures such as the Oral Health Impact Profile (Slade & Spencer, 1994). Most of the measures currently available use ordinal scale response formats (e.g. Likert scales) which attempt to capture frequency and severity of impacts across a range of conceptual domains.

9.3.3 Validity

A measure must be *valid* for the purpose for which it is being used, i.e. is it actually measuring what you are trying to measure? There are many types of validity, and the reader should consult Streiner and Norman (2003) for a more complete review of this topic. However, there are some measurement issues which are considered further here.

The questions in the measure must be relevant to the condition in question. *Generic* measures are designed for a range of conditions, and are not disease or context specific. In this situation, one may expect a certain degree of *item redundancy*, i.e. a high prevalence of items which are not relevant to the context (e.g. the oral cavity) or the condition (e.g. tooth loss impacts) of interest. There must be a sufficient number of items within the measure to capture consequences of disease. Furthermore, where item redundancy is significant, the measure may be prone to floor and/or ceiling effects. Generic measures can be used to measure health status across a range of populations, but the trade off comes with higher prevalence of

item redundancy. *Disease specific* measures are context specific, and potentially more useful when a specific condition is the focus of interest. It is not uncommon for researchers to use both generic and disease specific measures, but this increases the burden on potential respondents. The potential difficulty in using generic measures is illustrated in a randomized clinical trial of treatment provided to edentulous patients reported by Bouma *et al*. (1997). Three treatment modalities in the treatment of severely resorbed edentulous mandibles of edentate adults were compared, namely: (1) implant supported overdentures; (2) vestibuloplasty with conventional complete denture, and (3) conventional complete denture only. Quality of life outcomes were compared using the Hopkins Symptom Checklist, a Linear Analogue Self-Assessment of quality of life, the Groningen Activity Restriction Scale and the Psychological Well-Being Scale for denture patients. None of these measures is specifically designed to record oral health outcomes, and are considered *generic* health status measures. There were no significant differences between the groups at baseline. Following treatment, significant improvement was reported in each scale in all three groups. A noteworthy finding was that the degree of improvement was the same in all groups, suggesting that implant therapy was no more effective than conventional approaches in improving quality of life. This could be explained by the use of measures which were not oral specific to assess outcomes of oral health change.

Discriminant validity is of relevance in the context of cross-sectional population studies. A measure with good discriminant validity properties will be characterized by large and stable between-subject/group variation. An example from a study reported by Allen and McMillan (2003). In this study, two measures were used to assess oral health related quality of life in edentulous patients. In one group (IG), these patients requested implant retained prostheses, whereas the second group (DG) simply requested replacement of their existing dentures. Although clinical presentation was the same, i.e. they were edentate, motivation for seeking treatment was different between the groups. A measure with good discriminant validity properties, and appropriate content validity (i.e. sufficient number of questions/items relevant to the problem of interest) should be capable of detecting differences in the consequences of disease. At baseline, the measures used were the SF-36 (Ware & Sherbourne, 1992), a widely used general health status measure, and the Oral Health Impact Profile (OHIP) (Slade & Spencer, 1994). The SF-36 contains 35 items divided into 8 conceptual domains, and a single self reported health transition statement. Domain scores are calculated on a 0–100 scale. The OHIP contains seven domains, and domain scores are calculated by summing item impact prevalence response codes (ranging from 0 = never to 4 = very often). As might be expected, the authors reported that the oral specific measure detected between group differences, whereas the SF-36 did not.

9.3.4 Responsiveness

The ability to detect change is critical in determining the impact of clinical interventions. If a researcher can demonstrate that one form of intervention is markedly

more effective than another in its impact on quality of life, then this may have importance in a public health funding context. Unfortunately, not all measures can be assumed to have good *responsiveness* properties, i.e. the ability to measure change over time in clinical trials or longitudinal studies.

In general, the higher the proportion of 'high prevalence' items (i.e. items with a high frequency), the better the measure is likely to be at detecting change. For example, should the researcher wish to measure the impact of tooth loss on chewing, it would be important to have a number of questions related to chewing difficulty in order to capture this impact.

Measurement of change can be complicated by a number of phenomena. In essence, quality of life is a dynamic construct, and prone to influence by changing life circumstances and environmental issues as well as biomedical factors. The phenomena of *Response Shift* and *Implicit Theory of Change* have been described elsewhere (Streiner & Norman, 2003), and should be considered when interpreting change scores. According to the theory of *Response Shift* (Schwartz & Sprangers, 1999) a follow-up response may be influenced by new information not available at the time of the initial response. This theory is used to explain phenomena such as differences between objective/professionally assessed health status and patient rated health status. *Implicit Theory of Change* (Ross, 1989) cannot recollect their previous health state, and tries to work out their perceived health status relative to their current health status. The measurement challenges presented by both theories are complicated, and it seems likely in both theories that retrospective judgements of previous states are likely to be biased.

Health-related quality of life measures are frequently reported as mean summary scores either for subscales/domains or for by creating a summary variable for the entire responses given. For most measures, 'higher' scores tend to be equated with poorer self-rated health status. Summary scores tend to be non-normally distributed, and are skewed to the left of the distribution plot. Accordingly, these data are subject to spontaneous within-subject change, random measurement error or both. This is particularly likely when chronic disease states are being assessed. The goal of a treatment intervention is to, in effect, result in a lowering of scores which in turn equates to an improvement in self-reported health status. However, by their very nature, these chronic conditions are capable of spontaneous improvement or deterioration without clinical intervention. Accordingly, great care must be taken to ensure that reported 'improvements' are, in fact, attributable to the clinical intervention and not spontaneous improvement unrelated to the intervention in question.

The simplest approach taken to measure change is to subtract the post-treatment scores from the pre-treatment scores. This approach assumes that impacts will be similar for all subjects under study, but this is unlikely to be the case. There is a possibility of *regression to the mean* (see Chapters 6 and 11 for further detail) when within group pre- and post-intervention comparisons are made. Extremely skewed (left or right) scores are strongly influenced by random error effects, and may change spontaneously without clinical or therapeutic intervention. In practical terms, it is possible to ascribe a 'benefit' to a clinical intervention, which may in fact not be due to the intervention at all and merely caused by regression to the

mean. This issue has been discussed by Worthington (1998) and Slade (1998), and both recommend that the value of clinical intervention can only be determined by making between group comparisons using an appropriate control group.

9.3.5 Practical utility – use of shortened versions of health status measures

In terms of practical utility, the more items in a questionnaire, the more cumbersome it becomes to use. In the context of a clinical trial, it is simply not practical to use a measure containing more than 20 or so items. It may also be impractical to include large questionnaires in cross-sectional population surveys. However, as items are removed from a questionnaire to improve its utility, the measure loses precision and this in turn impacts upon its measurement properties. It is unwise to attempt to reduce the number of items in a health status measure by choosing items in an ad hoc fashion. A variety of statistical methods have been used to reduce the number of items in health status measures. These include factor analysis, internal reliability analysis and least squares regression analysis.

A potential difficulty with using statistical methods to produce sub-sets of items is that potentially important items can be removed. Coste *et al*. (1995) have cautioned that part-whole correlation effects is likely when forward stepwise regression procedures (using the long form as the dependent variable) are used to enter items into a short-form measure. An example of this is the shortened version of the Oral Health Impact Profile (OHIP-14), which is a subset of the original 49 item measure. Using the OHIP summary score as the dependent variable, least squares regression analysis was used to produce the 14 items, two items for each of the seven conceptual domains in the measure. However, this subset does not contain items related to chewing, which is of concern to people with missing teeth. A potential problem with this in clinical trials is that the measure is dominated by items which are not likely to change following clinical intervention, i.e. subject to 'floor' and 'ceiling' effects. In turn, this may mean that actual change may be masked by not having items in the measure likely to be sensitive to change.

Juniper *et al*. (1997) have used a clinical item impact method for generating items. In this method, both frequency and importance of impacts for each statement in a health status measure are calculated. The 'frequency' of the impact is defined as the proportion of the sample who report experiencing this impact. The 'importance' of each item is determined by calculating the mean response for each item. An 'item impact' is calculated as follows:

$$Item\ Impact = Frequency\ x\ Importance$$

Item impact scores are then ranked, and items with higher scores are deemed to be more relevant or important to patients than lower scores. The researcher can then select items of relevance to the clinical context when, for example, planning an evaluation of a clinical intervention. Allen and Locker (2002) have used this approach to produce a subset of OHIP items (OHIP-EDENT) which can be used

to evaluate clinical outcomes for edentulous patients. The subset of items produced using the Item Impact method produced some overlap with the OHIP-14, but critically, also had items relevant to chewing and eating difficulty. Using the full 49 item OHIP as a *'gold standard'*, measurement properties of the OHIP-14 and the OHIP-EDENT were compared in a clinical trial of edentulous patients. The data indicated that OHIP-14 and OHIP-EDENT had similar discriminant validity properties, but the OHIP-EDENT had superior sensitivity to change. As described earlier, this study highlights two important principles, namely:

- the need to decide whether you wish to use a measure to detect between subject/group variability (discriminant measure), or, change following a clinical intervention (evaluative measure);
- the need to have sufficient item representation within the measure to tap into problems experienced by the respondent.

9.3.6 Weighting

Item weighting refers to a process whereby it is possible to allow the relative importance of events to be described. However, it also complicates the scoring of individual items, and an overall analysis of data. In practice, it appears that measurement properties are not significantly improved by weighting of items (Jenkinson, 1991). A comparison of measurement properties of weighted and unweighted versions of the OHIP has been made by Allen and Locker (1997). They found that discriminant validity of the OHIP was improved slightly by using the weighted scoring system, but questioned whether this improvement made use of a complex scoring system worthwhile. They suggested that the unweighted methods of scoring the OHIP were adequate for use in clinical settings. This finding was also confirmed by McGrath and Bedi (2004).

9.3.7 Reference periods

On a practical level, it is important to frame questions in a relevant time period, i.e. a 'reference period'. For example, the Oral Health Impact Profile uses a question format 'In the past *year*, have you had ... because of problems with your teeth, mouth or dentures?' There is little standardization of quality of life measure reference periods, and some questionnaires use reference periods of weeks, whereas others use 6 or even 12 month reference periods. As described earlier, it is important to have considered this issue in advance of collecting data, and to have decided on an appropriate reference period. If the aim is to collect data pertaining to items likely to fluctuate, e.g. painful conditions, then a short reference period of weeks would be recommended. A short reference period is also likely to be appropriate in clinical trials when measuring the immediate impact of a clinical intervention. On the other hand, some dimensions of quality of life may be expected to show a greater degree of temporal stability, e.g. psychosocial factors. If these items are the likely focus of interest, then a longer reference period of 6 or 12 months would

be preferable. This is likely to be the case in national population health surveys, and between country comparisons would be greatly facilitated with a standardized approach to reference periods. In a recent paper on this issue, Sutinen *et al*. (2007) recommended using 12 month reference periods for national surveys.

9.4 Summary

In order to improve insight into the impact of dental disease, it is vital to obtain good quality data from health surveys. This will involve collection of clinical and subjective data, and care needs to be taken to collect these data in a systematic manner. Furthermore, stringent methods need to be put in place to deal with consistency issues and for handling missing data. Finally, it must be recognized that measurement property requirements for collecting subjective data vary between cross-sectional population surveys and clinical trials. Failure to take these issues into account may lead to spurious interpretation of data collected in health surveys and clinical trials.

References

Allen, P.F. & Locker, D (1997) Do weights matter? An assessment using the Oral Health Impact Profile. *Community Dental Health* **14**: 133–8.

Allen, P.F. & Locker, D. (2002) A modified short version of the Oral Health Impact Profile for assessing health related quality of life in edentulous adults. *International Journal of Prosthodontics* **15**: 446–50.

Allen, P.F. & McMillan, A.S. (2003) A longitudinal study of quality of life outcomes in older adults requesting implant prostheses and complete removable dentures. *Clinical Oral Implants Research* **14**: 173–9.

Altman D (1999) *Practical Statistics for Medical Research*. Chapman & Hall/CRC.

Bouma, J., Boerrigter, L., Van Oort, R.P., van Sonderen, E., & Boering, G. (1997). Psychological effects of implant-retained overdentures. *International Journal of Oral and Maxillofacial Implants* **12**: 515–22.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* **20**: 37–46.

Cohen, J. (1977) *Statistical Power Analysis for the Behavioural Sciences*. New York: Academic Press.

Coste, J., Fermanian, J., & Venot, A. (1995) Methodological and statistical problems in the construction of composite measurement scales. A survey of six medical and epidemiological journals. *Statistics and Medicine* **14**: 331–45.

Dworkin, S.L. & LeResche, L. (1992) Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. *J Craniomandib Disord*. **6**: 301–55.

Fitzpatrick, R., Fletcher, A., Gore, D., Spiegelhalter, D., & Cox, D. (1992) Quality of life measures in health care. I: Application and issues in assessment. *British Medical Journal* **305**: 1074–7.

Guyatt, G.H., Feeny, D.H., & Patrick, D.L. (1993) Measuring health-related quality of life. *Ann Int Med* **118**: 622–9.

Ismail, A.I., Sohn, W, Tellez, M., *et al*. (2007) The International Caries Detection and Assessment System (ICDAS): an integrated system for measuring dental caries. *Community Dentistry and Oral Epidemiology* **35**: 170–8.

Jenkinson, C. (1991) Why are we weighting? A critical examination of the use of item weights in a health status measure. *Soc Sci Med* **32**: 1413–16.

Juniper, E.F., Guyatt, G.H., Streiner, D.L., & King, D.R. (1997) Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epidemiol* **50**: 233–8.

Landis, J.R. & Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**: 159–74.

Lesaffre, E., Mwalili, S.M., & Declerck, D. (2004) Analysis of caries experience taking inter-observer bias and variability into account. *Journal of Dental Research* **83**: 951–5.

Locker, D. (1988) Measuring oral health: a conceptual framework. *Community Dental Health* **5**: 3–18.

McGrath, C. & Bedi, R. (2004) Why are we 'weighting'? An assessment of a self-weighting approach to measuring oral health related quality of life. *Comm Dent Oral Epidemiol* **32**: 19–24.

Ross, M. (1989) Relation of implicit theories to the construction of personal histories. *Psychological Review* **96**: 341–57.

Schwartz, C.E. & Sprangers, M.A.G. (1999) Methodological approaches for assessing response shift in longitudinal health related quality of life research. *Social Science and Medicine* **48**: 1531–48.

Slade, G.D. (1997) Derivation and validation of a short-form oral health impact profile. *Community Dentistry and Oral Epidemiology* **25**: 284–90.

Slade, G.D. (1998) Assessing change in quality of life using the Oral Health Impact Profile. *Comm Dent Oral Epidemiol* **26**: 52–61.

Slade, G.D. & Spencer, A.J. (1994) Development and evaluation of the Oral Health Impact Profile. *Community Dental Health* **11**: 3–11.

Slevin, M.L., Plant, H. & Lynch, D. *et al*. (1988) Who should measure quality of life, the doctor or the patient? *British Journal of Cancer* **57**: 109–12.

Streiner, D.L. & Norman, G.R. (2003) *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford: Oxford University Press.

Sutinen, S., Lahti, S., & Nuttall, N., *et al*. (2007) Effect of a 1-month vs a 12-month reference period on responses to the 14-item Oral Health Impact Profile. *Eur J Oral Sciences* **115**: 246–9.

Ware, J.E. & Sherbourne, C.D. (1992) The MOS 36- item short-form health survey (SF36). I. Conceptual framework and item selection. *Medical Care* **30**: 473–83.

Worthington, H.V. (1998) Statistical aspects of measuring change in oral health status of older adults. *Comm Dent Oral Epidemiol* **26**: 48–51.

Part III

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

10

Start with the basics

Manal A. Awad, Nico Nagelkerke and Emmanuel Lesaffre

10.1 Introduction

This chapter introduces statistical methods that allow oral health researchers to describe the information collected in a sample, and to see how much their findings can be generalized to the population as a whole. Different types of data collected in oral health literature are introduced and we demonstrate how to organize and present summaries of the gathered data. Inferential statistics are also introduced. Throughout this chapter, examples from the oral health literature will be used to explain different statistical approaches. The main focus of this chapter is on the use and interpretation of the basic statistical concepts necessary to perform oral research. Mathematical expressions and derivations will be kept to a minimum.

Symbols and formulas for descriptive statistics vary depending on whether one is describing a sample or a population. Characteristics of a population are referred to as parameters while characteristics of samples are called statistics. In order to distinguish between them, Greek letters are used to denote parameters; a parameter with a 'hat' denotes an estimate of a parameter obtained from the gathered data. Roman letters are used to denote statistics.

Statistical analyzes of oral health data are challenging because of the hierarchical nature of many oral health data, e.g. mouth, teeth, surfaces. However, we defer the statistical implications of this special structure of the data to later chapters.

In this chapter we deal with the classical statistical approach. An alternative approach is described in Chapter 18.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

10.2 Variables and scale of measurement

A variable is a measured characteristic that varies among persons, events or objects being studied. The scale of measurement is the scale on which a characteristic is measured, which has implications on how information is displayed, summarized and analyzed. Note that the statistical analyzes also depend on other factors besides the scale, such as the research question. The following scales of measurement occur most often in oral health research: (1) categorical: nominal, ordinal and binary; and (2) numerical: continuous and discrete.

Nominal scales are used for the simplest level of measurement, when data values fit into categories. Nominal scales have no order and are generally category labels that have been assigned to classify items or information. Variables with categories such as profession, and place of birth are nominal scales which could be assigned values. For example, profession can be classified as 1 = unemployed, 2 = blue-collar worker, 3 = white-collar worker, etc. A special case of a nominal variable is when there are only two classes, in that case the variable is called dichotomous or binary, e.g. gender is a binary variable with, say 1 = male and 2 = female. When the categories have an inherent order, then observations are said to be measured on an ordinal scale. For example, the severity of dental fluorosis using Thylstrup-Fejerskov Index in which a score of '0' indicates no fluorosis and a score of '9' indicates severe enamel damage due to fluorosis is an example of an ordinal scale. An important characteristic of an ordinal scale is that differences between adjacent categories are not equal throughout the scale. To illustrate this, the difference in degree of dental fluorosis between a score of 7 and a score of 8 is probably not the same in magnitude as the difference between 1 and 2.

A numerical scale is a scale in which the differences between numbers have a meaning, because they measure a particular quantity. There are two types of numerical scales, continuous scales and discrete scales. Continuous scales have values on a continuum, such as age, mandibular bone height, length of survival of implants. Discrete scales have values equal to integers, such as number of decayed surfaces, number of oral implants placed in an edentulous mandible.

10.3 Measures of central tendency (location)

10.3.1 Mean

The mean or average describes the center of gravity of a frequency distribution. The mean of a sample $x_1, \ldots, x_n$ is symbolically represented by $\bar{x}$ read as 'x bar'. To compute $\bar{x}$, the values x_i (x-value of the ith individual) are added up and divided by the sample size, n. This is expressed with the following expression(s):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{10.1}$$

where the upper case Greek letter sigma $\left(\sum\right)$ means 'the sum of', and $\sum_{i=1}^{n}$ means that the sum is taken from the 1st to the nth observation. A shorter notation is $\sum_i$.

The data from a randomized controlled clinical trial (RCT) comparing mandibular two-implant overdentures and conventional dentures in 102 edentulous patients are used for various illustrations. In this study, referred to in the remainder of the chapter as the *Two-Implant Overdenture (2-IO) Study*, baseline data were collected on patients' ratings of general satisfaction with their current mandibular prostheses on a 100 mm Visual Analogue Scales (VAS), in which low ratings indicate less satisfaction, see Awad *et al*. (2003). The mean age of these patients is equal to 50.91 yrs and their average rating is equal to 38.5 mm.

The mean is used when the values can be added, i.e. when the characteristics are measured on a numerical scale. The mean is, however, not appropriate for nominal data. For example, a study that compares different types of implants assumes a code '1' for an implant-retained overdenture on 2 implants with ball attachments, while a code '2' is used for an implant-retained overdenture on 2 implants with a single egg-shaped bar and an implant-retained overdenture on 4 implants with a triple bar is coded as '3'. Assuming that an average of 2.3 is obtained, then this value is meaningless since it depends on the particular choice of codes. Calculating a mean for ordinal data is often done in practice, but it is not immediately clear whether it is appropriate since its meaning depends again on the actual coding of the ordinal variable. For example for the ordinal variable 'very bad', 'bad', 'ok', 'good', 'super' one could give numerical scores 0 to 4, respectively. However, scoring these ordinal classes as 0, 1, 5, 10, 30 still complies with the ordinal character of the data, yet the mean will be drastically different from the first mean.

10.3.2 Median, quartiles and percentiles

The median is the middle value of the sample $x_1, \ldots, x_n$. It is the value below which 50 % of the observations are smaller and above which 50 % are larger. In the 2-IO Study, the median age is equal to 51 yrs and the median satisfaction score is 27 mm. A useful feature of the median is that it is not sensitive to extreme values in a data set. For instance, if for one individual a typo was made when entering the satisfaction data into the computer, e.g. instead of 72 one typed 7722 then the mean rating changes dramatically while the median stays at 27. The median is also called the 50 percentile (50 %-ile) or the 2nd quartile and denoted in this respect by Q_2. One can calculate in a similar manner the 25 %-ile (1st quartile or Q_1) and the 75 %-ile (3rd quartile or Q_3). In general, the $m\% - ile$, with m an integer corresponds to the value for which m % of the observations are smaller. The first and third quartile of the satisfaction ratings in the 2-IO Study are 12 mm, 67.5 mm respectively. Note when quartiles and percentiles are computed from small data sets, the reported value will most often not satisfy exactly the definition. Further, the reported value will vary also with the chosen software.

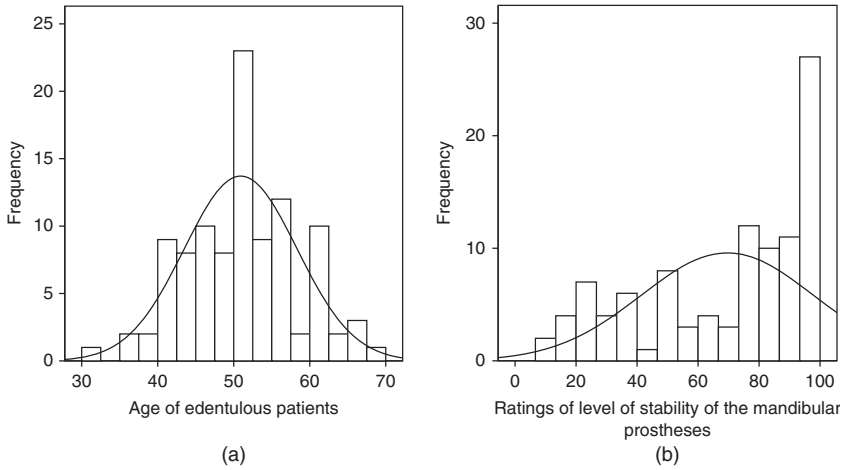


Figure 10.1 2-IO Study: shapes of distributions (displayed as histograms) and approximating normal curves. (a) symmetric and (b) negatively skewed.

10.3.3 Comparison of measures of central tendency

Which measure of central tendency is best to use in practice, the mean or the median? Two factors should be considered to make the choice: (a) the scale of measurement and (b) the shape of the distribution. For numeric variables, a distribution is said to be symmetrical when the shape is (roughly) the same on both sides of the mean. For an example, see Figure 10.1a. If there is a tail on the right hand side of the distribution, then it is said that the distribution is skewed to the right or positively skewed. In that case the mean will be larger than the median. While if there is a tail on the left hand side, then the distribution is skewed to the left or negatively skewed, see Figure 10.1b and then the median will be larger than the mean.

For numerical data, the mean characterizes the distribution as a central measure when the distribution is symmetric, but also the median is a central measure. When the distribution is (severely) skewed, then the mean looses its property as a central measure in contrast to the median which always corresponds to the value for which 50 % of the observations are left to it. Hence for a skewed distribution the median is the best choice to characterize the distribution. For ordinal data, the best choice is the median, but as noted above, the mean is also used in practice at one own's risk.

It is important to note that even for skewed distributions the mean might be used for comparative purposes (e.g. income), even if it is not the best measure for centrality.

10.4 Measures of variability (spread)

It is possible for two data sets to have the same average but showing different variability. When the values in a data set are close to each other then the data

show low variability, i.e. they are homogeneous. In the other case, they have high variability, i.e. they are heterogeneous. Two types of measures of variability are discussed here, the standard deviation (SD) and the inter-quartile range (IQR). There is confusion in the literature between the standard deviation and a measure called the standard error of the mean (SEM) or also shorter the standard error (SE). The latter is not a measure of variability of the original data, but rather of the sample mean and will be treated in Section 10.7.4.

10.4.1 Standard deviation

The standard deviation is the most commonly used measure of variability and is based on $x_i - \bar{x}$ $(i = 1, \ldots, n)$ which are the deviations of the observations from the mean. A value x_i that is remote from the mean will produce a large negative or positive value for this difference. Many large deviations imply that there is a lot of variability of the individual data around the mean. A measure that expresses the overall variability is the variance s^2 given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{10.2}$$

Note that in (10.2) the variance is a summation of squared deviations; this to avoid that positive and negative deviations cancel out. Also, it is (almost) the averaged (because of denominator $(n-1)$) sum of squared deviations such that a larger sized sample will not have systematically a larger variance. The standard deviation is defined as the square root of s^2, i.e. s or also denoted SD. For the 2-IO Study, the standard deviation of age is equal to 7.43 yrs and 31.9 mm is the standard deviation of the ratings.

The SD is useful in comparing the variability around the mean between different groups. It gets a quantitative meaning for the variability around the mean when the distribution is symmetric (read: 'normal' or 'Gaussian' distribution, see below). Indeed, when the data have a symmetric (normal) distribution, 68 % of the values in the data set are within 1 SD of the mean and 95 % of the values are within 2 SD from the mean. This justifies the classical notation of mean±SD, often reported in the oral health and medical literature. However, if the data don't have a symmetric (normal) distribution then this notation might be close to meaningless, i.e. often mean-SD (when SD is relatively large what happens in skewed distributions) extends beyond zero and the interval [mean-SD, mean+SD] contains much more than 68 % of the observations! This was observed for the satisfaction ratings in the 2-IO Study where the standard deviation is almost equal to the mean.

Knowledge of the SD might be useful on a clinical level since it allows for the clinicians to know if a particular observation is 'normal'. For example, it is well known that saliva plays a significant role in the maintenance of oral health, and many studies have reported that a low salivary flow rate is a risk factor for dental caries, periodontal diseases and candidiasis. Suppose as a dentist you have a 25 year old female patient with unstimulated whole saliva flow rate (UWSFR)

of 0.21 ml/min. You remember that studies reported that the average saliva flow in normal (read: healthy) individuals is around 0.35 ml/min. You would like to know if the patient with a UWSFR of 0.21 ml/min is within the normal range or not. If UWSFR has the correct type of distribution, then knowledge of the standard deviation allows you to judge this. Suppose SD = 0.12 ml/min, then plus or minus two SDs around the mean results in values that range from 0.11 and 0.59 and therefore the UWSFR of this patient is within the 'normal' range.

10.4.2 Inter-quartile range

The standard deviation receives a quantitative meaning only when the data have a normal distribution. A measure of variability that does not make this assumption is the interquartile range IQR, which is simply the 75 %-ile minus the 25 %-ile and expresses the length of the interval where 50 % of the middle values of the sample are located. This measure has always a clear quantitative interpretation irrespective of the distribution of the data.

10.5 Graphical display of data

Graphs are essential in the research and the reporting stage of oral health research. The type of graph depends on the kind of data, as will be seen below. We will restrict ourselves in this chapter to graphs representing characteristics of each variable separately, i.e. univariate graphs.

10.5.1 Graphs for nominal and ordinal data

A bar chart is typically shown for categorical data. A bar chart consists of bars with base located around the value on the X-axis that defines the class. The height of the bar is equal to the frequency, proportion or percentage of individuals with that class value. A pie chart presents proportions or percentages as pieces of a pie.

10.5.2 Graphs for numerical data

For numerical data, a bar chart is not an appropriate graphical tool when there are too many different values. In that case a histogram is preferable. A histogram shows the frequency (or proportion/percentage) of measurements in classes defined by the user (and/or statistical software) and therefore provides an idea of the shape of the data. Figures 10.1a,b are histograms from which we concluded the a (symmetry) of the distribution of the gathered data. Note that the appearance of the histogram depends on the choice of classes.

Another type of plot, the error-bar plot, is typically used for the graphical comparison of groups. See Figure 10.2a for the error-bar plot on the satisfaction ratings obtained at baseline in the 2-IO Study split according to gender. For each group the mean is indicated with a short horizontal line while two vertical bars extend from mean-SD to mean+SD. Note that the error-bar plot implicitly assumes

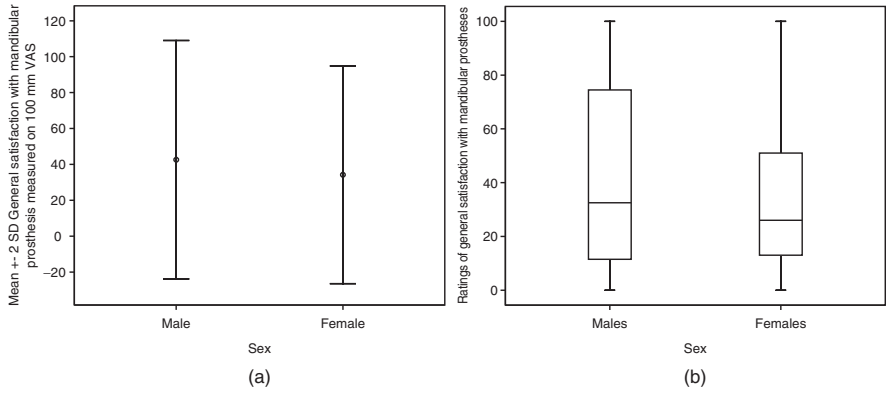


Figure 10.2 2-IO Study: (a) error-bar plot (left) and (b) box(-whisker) plot (right) of satisfaction ratings by gender.

that the distribution of the numerical value is symmetric in each group. A related, and more popular plot in oral health and medical journals appears like a bar-chart with height equal to the mean and on top a bar extending to mean+SD.

A box plot or a box- and whisker plot is a graphical display of data based on quartiles. The box plot of ratings of general satisfaction in the 2-IO Study is shown in Figure 10.2b, for males and females separately. A box is drawn with the top at the 3rd quartile and the bottom at the 1st quartile. The length of the box is the visual representation of the IQR. The median is indicated with a horizontal line. Straight lines or whiskers extend below and above the box, but the length of these whiskers very much depends on the statistical software. In Figure 10.2 they extend to the smallest (largest) observed value that isn't an outlier. Here an 'outlier' was defined as a value between 1.5 and 3 box lengths from the upper or lower edge of the box. The definition of an outlier is very much software dependent, and clearly does not allow the user to omit this aberrant value from the analysis later on.

The distribution of the data can be inferred from a box-plot. When the median lies asymmetric in the central box and/or the whiskers are unequal in length, the distribution must be skewed as for the satisfaction ratings.

10.6 Understanding probability

Probability lies at the heart of statistical theory. The probability of an event A, $P(A)$ is a positive value between 0 and 1 that measures the chance that A occurs in the following way. Assume that an experiment can be repeated many times, with each repetition called a trial. Assume that event A can result from each trial, then $P(A)$ equals the number of times that A occurs divided by the total number of trials. If $P(A)$ is equal to zero then A cannot occur, while A is sure to occur when $P(A) = 1$.

10.6.1 Random variables and probability distribution

The value that a variable takes, varies from one individual to another. A discrete random variable X is a quantity that can take any one of a set of values with a given probability. A probability distribution shows the probabilities of all possible values of the random variable. When the variable X is continuous, such as age, there are an infinite number of possible values and then probability statements pertain to events like 'age is in-between 30 and 50'. The value that X can take is indicated by the lower case symbol x. The values $P(X = x)$ for x can be summarized in a frequency distribution called a (discrete) probability distribution. For a continuous variable one speaks of a probability density (function). Both the probability distribution and the density are theoretical functions expressed mathematically as a function of the values x that X can take. They classically depend on parameters, say α, β, ... and are denoted by e.g. $f(x; \alpha, \beta, \dots)$. The most prominent example is the normal distribution introduced below in which there are two parameters, μ and σ^2. In general, the probability distribution represents the distribution of a measurement taken on a population of subjects. In this sense it has a population mean and variance, they are both determined from $f(x; \alpha, \beta, \dots)$. For instance, the mean for a continuous random variable X is calculated as follows

$$\mu = \int x f(x; \alpha, \beta, \dots)\, dx, \tag{10.3}$$

whereby $f(x; \alpha, \beta, \dots)$ is the density of X and '$\int$' represents an integral. Note that expression (10.3) is the limit version of expression (10.1), whereby the sum turns into an integral.

Several theoretical probability distributions are important in statistics, such as the binomial distribution, the Poisson distribution and the normal or Gaussian distribution.

10.6.2 The binomial and Poisson distribution

The binomial distribution is the distribution of the random variable X which is the number of 'successes' in n experiments. If the probability of success in one experiment is equal to π, then the probability of $X = k$ is

$$P(X = k) = \binom{m}{k} \pi^k (1 - \pi)^{(m-k)} \qquad (k = 0, \dots, m), \tag{10.4}$$

where $\binom{m}{k} = \frac{m!}{k!(m-k)!}$ is called the *binomial coefficient* and $k! = k \times (k - 1) \times (k - 2) \times \cdots \times 2 \times 1$. By convention, $0! = 1$. For example, suppose that the probability of demonstrating caries experience (CE), at level d_3, for seven year old children is $\pi = 0.55$ then the probability that $X = 0, 1, 2 \dots$ deciduous teeth out of $n = 20$ show caries experience at that age is given by Figure 10.4.

A related probability distribution is the Poisson distribution. The following example illustrates its usefulness. There is currently an increasing awareness of the medical errors causing serious adverse events to the patient that happen in

especially large hospitals. Suppose that, every year on average five severe medical errors happen in a dental unit of a large hospital. Assuming that this average applies to all large hospitals in the world and that the errors occur independently of each other, we can calculate the probability that k severe medical errors will occur in the future. Namely, the Poisson distribution tells us that

$$P(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} \qquad (k = 0, \ldots,), \tag{10.5}$$

where X represents the number of medical errors on a yearly basis in the dental unit of a large hospital, λ represents the average number of severe medical errors in one year (hence equal to five in our example) and e is a constant equal to about 2.718282. Note that k is now unlimited, at least in theory, so that $\sum_{k=0}^{\infty} P(X = k) = 1$. The Poisson distribution corresponding to expression (10.5) is denoted Pois(λ) and its graphical representation for $\lambda = 5$ is given in Figure 10.3.

The Poisson and the binomial distribution are related, i.e. for $\pi \times m$ small the binomial $P(X = k)$ is close to the corresponding Poisson probability with $\lambda = \pi \times m$. Further, both the binomial and the Poisson distribution are based on the assumption that the events occur independently. If this assumption is violated then these models will not necessarily give a correct description of the probabilities that occur in the population. For instance, the dmft-index represents a sum of 'successes' with $m = 20$, the total number of deciduous teeth. In Chapter 20 the Signal-Tandmobiel® study is discussed in detail. Here we limit ourselves to saying that the ..., the Signal-Tandmobiel® study is a longitudinal study in which children were examined annually with respect to their oral health condition, dietary habits, etc. from the first year in primary school onwards.

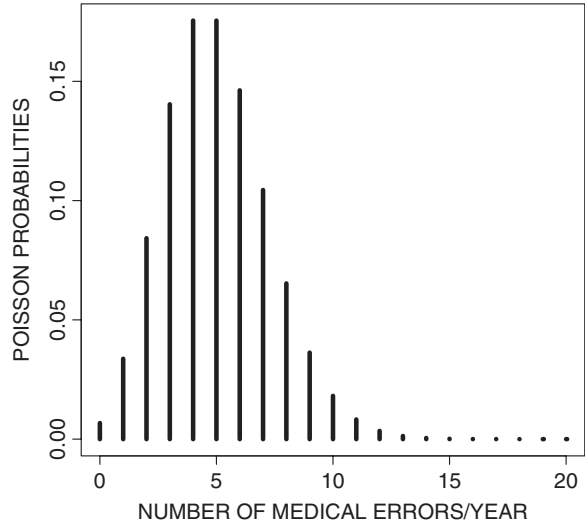


Figure 10.3 Poisson probability distribution with mean equal to 5 (Pois(5)).

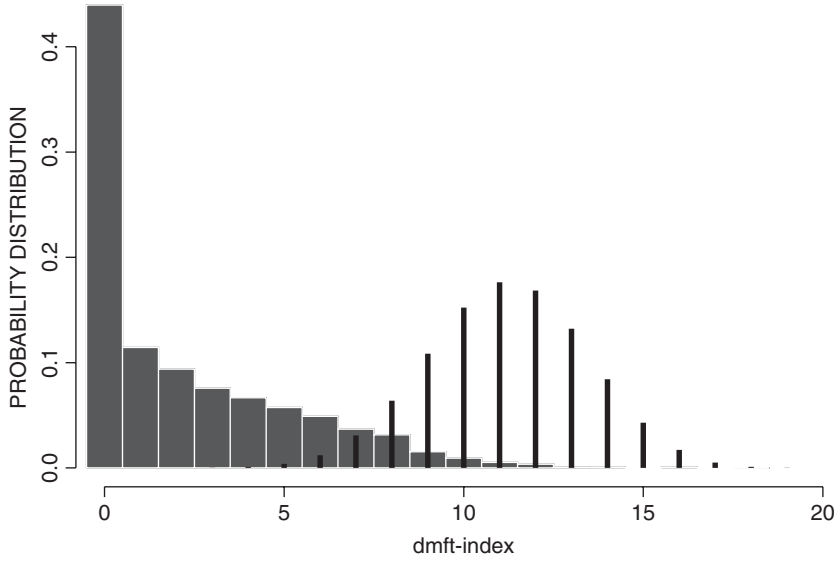


Figure 10.4 Comparison of Signal-Tandmobiel® probability distribution of dmft-index (left histogram) to binomial distribution based 20 experiments with probability of success' = 0.55, i.e. Bin(20,0.55).

Act now as if the sample of 4351 children is an infinite population, i.e. assume that its histogram expressing proportions is in fact the description of the probabilities of X (dmft-index) in a population. If CE on teeth occurs independently, the dmft-index should have a binomial distribution Bin(20, 0.55) since there are 20 deciduous teeth and 0.55 is the average dmft-index. Figure 10.4 compares the Signal-Tandmobiel® probability distribution with Bin(20, 0.55). Clearly, the Signal-Tandmobiel® probability distribution is not well represented by the binomial distribution. The reason for this is that the probability of caries experience varies with tooth and that in the same child caries on one tooth is related to caries on another tooth.

10.6.3 The normal (Gaussian) distribution

The normal distribution occupies a central position in statistics. It is a continuous distribution which extends from $-$infinity to $+$infinity. It is also called the Gaussian distribution, because the mathematician Carl Friedrich Gauss extensively used it for analyzing astronomical data and defined the mathematical formula of the probability density function. It is often called the bell (shaped) curve because the graph looks like a bell. The distribution (density) is symmetric about its mean μ. The standard deviation of the distribution is σ. For a graphical display of the normal density, see Figure 10.5. The parameters μ and σ^2 determine completely the normal distribution and is therefore denoted by N(μ,σ^2). The normal curve can be used to calculate

the probability of getting certain outcomes when taking samples from a population. For instance, the probability that a normally distributed random variable X with mean μ and standard deviation σ lies between: (a) $\mu - \sigma$ and $\mu + \sigma$ is about 68 %; (b) $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is about 95 % and (c) $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$ is about 99 %.

We illustrate the use of the normal distribution by the age of the children of the Signal-Tandmobiel® study at entry. The average age is 7.1 years with a standard deviation of 0.41 years. We can now fit a normal distribution to the histogram of ages with the same mean and standard deviation, see Figure 10.5. Using the above properties of the normal distribution we could claim e.g. that 68 % of the ages lie between 6.67 and 7.49 years. This will be close to the truth when the normal distribution is a good fit to the data. In the sample we found 64 %, indicating a relatively good but not perfect fit of the data to the normal distribution. Clearly, using the properties of the normal distribution allows for an easy calculation of all such kind of percentages. This was also extensively used at the time when computers were not available.

A special case of the normal distribution is obtained by taking $\mu = 0$ and $\sigma = 1$ yielding the standard normal distribution, denoted N(0,1). Many statistics textbooks contain a table expressing the ordinate to the area under the curve (or

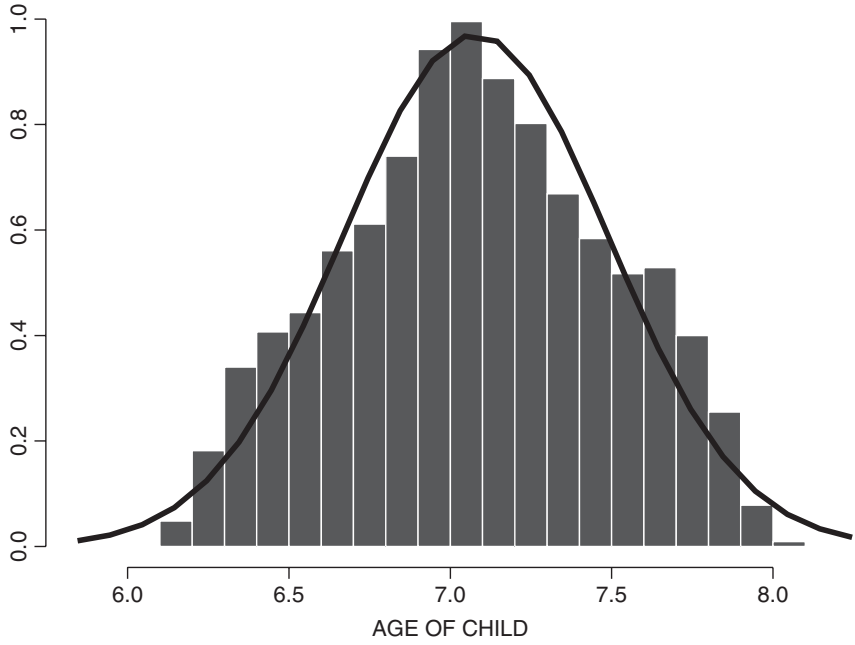


Figure 10.5 Signal-Tandmobiel® study: Histogram of age of children at first annual oral health examination and normal distribution N(7.1,0.41^2) (solid line) with same mean and same standard deviation as observed in sample.

equivalently the probability) left to it. For instance, the table would contain 0 next to 0.50 which is the area under the curve left to zero. In a similar manner, -1.96 and 1.96 correspond to 0.025 (2.5 %) and 0.975 (97.5 %), respectively.

From a normal random variable X, i.e. a random variable that has distribution $N(\mu, \sigma^2)$, a standard normal random variable Z can be derived as follows:

$$Z = \frac{X - \mu}{\sigma}, \tag{10.6}$$

i.e. the distribution of Z is N(0,1). An illustration of calculating the Z(-score) is the following. Suppose that the national mean DMFS for a particular population of adults between 40 and 50 years of age is 5 and the SD is 2. An individual from that population who has a DMFS score of 3, has $z = (3 - 5)/2 = -1$ as Z-score. The interpretation of this Z-score is simple. That is, for $Z = -1$, X must be equal to $\mu - \sigma$. Therefore, the Z-score tells you how many standard deviations from the mean a particular X-score is. Here, we will say that this person's DMFS score is by one standard deviation lower than the average DMFS of the population.

There are other continuous probability distributions used extensively in statistical inference (see sections below) such as: the t-distribution (also called Student's t-distribution), the χ^2-distribution, the F-distribution, etc.

Finally, the normal distribution is no doubt the most important distribution in statistics not because most data have a normal distribution but because of the Central Limit Theorem (see below), a quite remarkable result that is used in many of the statistical developments.

10.7 Principles of statistical inference

10.7.1 Population and samples

In health research, population in general refers to patients or other living organisms. In oral health research, the population can refer to individuals, teeth, surfaces of teeth or parts of the mouth. A population represents the entire group of subjects in whom a researcher is interested. However, it is quite costly to study an entire population, therefore data are collected on a subset of the population, called a sample, selected to be representative of the population. The data obtained from this sample are used to draw conclusions about the population.

10.7.2 Taking a representative sample

Probability theory assumes that researchers select independent subjects randomly from a particular population, this results in a simple random sample. Taking a simple random sample could be quite complicated in practice, because it often means that one visits and examines subjects physically in many different locations in a random order. That is why one has looked for other ways to make sampling more convenient in practice. This leads to cluster sampling. A cluster sample results

from a two-stage sampling process in which the population is divided into clusters and a subset of clusters are randomly selected. Clusters are commonly based on geographic areas or districts. This approach is usually used in epidemiologic studies, rather than in clinical studies. In oral health research, the cluster sampling approach has been used in national surveys of oral health and treatment needs of certain populations see e.g. Susin *et al*. (2006) and Naidu *et al*. (2006). The Signal-Tandmobiel® children have been obtained by a cluster sampling approach by first selecting schools and then selecting children from the schools. Moreover, the sampling was done in a stratified manner, i.e. cluster-sampling was executed in strata here the 15 combinations of 3 educational systems and 5 provinces of Flanders.

In practice, it could be sometimes difficult or time-consuming to obtain a random sample of subjects. Then researchers are inclined to select a convenience sample. For example, researchers can select patients from a single clinic or hospital and investigate some or all the patients with a clinical condition. This type of sampling results in a nonprobability sample, because the probability that a subject is selected is unknown and may reflect selection biases of the researchers conducting the study.

Classical statistical properties and statistical significance tests are based on the assumption of simple random sampling, but in case of cluster-random sampling the classical procedures can be adapted. More difficult is to appreciate the results based on a convenience sample and there are no clear ways to adjust the statistical procedures and hence it is not always clear what message a convenience sample is bringing us. Also important to note is that most of the developments in the remainder of this section assume that the random sample consists of independent subjects. Thus, when n individuals are sampled at random from a population, and on each subject 20 teeth are examined then it is not appropriate to act as if the total sample is $20 \times n$. Dedicated procedures are needed to deal with such dependent, also called correlated, data.

10.7.3 Point estimates

Researchers are often interested in the value of a parameter in the population. For example, let μ be the mean number of teeth extracted by age 60 or π be the proportion of 60 years old having dentures. Researchers estimate the value of the parameter using the data collected from the sample, referred to as sample statistic. In case of μ the sample statistic is the sample mean $\bar{x}$, also denoted $\hat{\mu}$ and for π it becomes the sample proportion $p = \hat{\pi}$. A sample statistic is called a point estimate of the parameter (a single value) as opposed to an interval estimate (see below) which takes a range of values.

10.7.4 The Central Limit Theorem and SEM

Let us assume that a study has been set up and a simple random sample of size n is taken from a population and that for each subject a measurement is taken.

Suppose that this measurement has in the population a mean equal to μ and a standard deviation equal to σ. An estimate of the population mean is the sample mean, i.e. $\bar{X}$ and an estimate of the population variance σ^2 is the sample variance s^2. The question is now how these estimates behave when repeated samples of size n are taken at random from the population. Thus, we suppose that a large number of sample means are obtained (in theory an infinite number). A classical result in statistics says that on average the sample means have a mean equal to the population mean of the individual data, namely μ. This property is referred to in statistics as unbiasedness, in fact one states that the sample mean is an unbiased estimate of the population mean. One can also show that the sample variance s^2 is an unbiased estimate of σ^2. In fact, the reason for having the denominator $(n-1)$ instead of n in expression (10.2) is that for the latter choice the estimate underestimates the true value and hence is biased. Further, the sample means vary from sample to sample but they do so with less variability than the original data. One can show that the variance of the sample means is equal to σ^2/n. These properties do not tell us yet how the sample mean varies around the true population mean. In other words, what is the distribution of the sample means upon repeated sampling? A most important result is given by the Central Limit Theorem (CLT) which states that for a large n (where the notion 'large' will be discussed below) the sample mean has a normal distribution $\mathrm{N}(\mu, \sigma^2/n)$ irrespective of the distribution of the original data. Note that the same result holds for all n when the original data have a normal distribution. For a large n, we can replace σ by s and the practical version of the CLT becomes that $\bar{X}$ has a $\mathrm{N}(\mu, s^2/n)$ distribution for a large n, or expressed in the Z-score:

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a N(0,1) distribution.

While at first glance this theoretical result is not of much practical use since one never knows the true mean in the population, the CLT is of utmost importance. Indeed, the CLT is the most important result in statistics and is used in many developments of classical statistics. The standard deviation of the sample mean $\sigma/\sqrt{n}$, estimated by $s/\sqrt{n}$ is called the standard error of the mean, denoted SEM or simply the standard error (SE).

There is a lot of confusion in the literature on the difference between SD and SE. To be clear, the SD expresses the variability of the individual data in the population while SE expresses the variability of the sample mean under repeated sampling. For example, assuming a population of adults between the ages of 25–40 have a mean of decayed, missing and filled surfaces (DMFS) equal to 3.6. A sample of 30 individuals was randomly selected from this population and the mean DMFS was found to be 3.7. For another random sample of 30 adults from the same population the mean was found to be 2.8. We can do this many times, each time we draw a random sample of 30 individuals. These sample means would form a normal distribution with the true mean as mean and with variance 30 times smaller than the original variance. Clearly, if we increase the sample size to, say 3000, then the majority of sample means will lie very close to the true mean. This is a deductive

result, i.e. we infer from the population to future samples. This result can however also be used for inductive reasoning, i.e. from the sample to the population as seen in next section.

10.7.5 Confidence intervals

It is important how precise the sample estimate estimates the population estimate. To express this, one uses the confidence interval. Suppose that the sample means have a normal distribution $N(\mu, \sigma^2/n)$ and assume that n is large. Then we know that under repeated sampling the probability that $\bar{X}$ lies between $\mu - 1.96\sigma/\sqrt{n}$ and $\mu + 1.96\sigma/\sqrt{n}$ is 0.95. This deductive reasoning gives also an inductive result after a simple reordering of μ and $\bar{X}$. Namely, suppose a study delivers an observed mean equal to $\bar{x}$ then the probability that μ is contained in $[\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n}]$ is 0.95. In practice the 95 % CI is estimated by $[\bar{x} - 1.96\, s/\sqrt{n}, \bar{x} + 1.96\, s/\sqrt{n}]$. This interval is called the 95 % confidence interval (CI) and it gives an idea of the uncertainty we have about the true population mean. The adjective '95 %' means that under repeated sampling the varying 95 % CI will contain the true mean with probability 0.95. Loosely speaking one states that for about 95 out of 100 studies the 95 % CI will contain the true population mean.

To illustrate the calculations, suppose that 100 edentulous patients who received mandibular 2-implant supported prostheses were asked to rate their level of satisfaction on a 100 mm Visual Analogue Scale. The sample mean of this rating was 75 mm, which is the point estimate for the unknown true value of the mean satisfaction rating of all edentulous patients in the same age group from this population, i.e. $\hat{\mu} = 0.75$. However, a single number does not give an indication of the uncertainty we have about the true mean. Confidence intervals allow us to quantify this uncertainty and estimate the range of values in which the true population mean would lie. In this example the 95 % CI was calculated as follows. As a first step we need to calculate the standard error of the mean (SE): $SE = s/\sqrt{n}$, here equal to $\text{SE} = 15/\sqrt{100} = 1.5$. Then the 95 % CI is $[75 - 1.96 \times 1.5, 75 + 1.96 \times 1.5]$ equal to [72.06,77.94]. The practical interpretation of the 95 % CI is as follows: we are 95 % confident that the true mean level of satisfaction in edentulous 60–75 years old patients who receive mandibular 2-implant supported prostheses is between 72.06 and 77.94. Thus the 95 % expresses our uncertainty on the true parameter. This is, however, a Bayesian interpretation of the 95 % CI, see Chapter 18.

Note that if we wish to calculate the 99 % CI, then 1.96 is replaced in the expression of the 95 % CI by 2.58. More important to note is that the width of the confidence interval depends on SE. The smaller SE, the narrower the CI and the less uncertainty we have about the true population mean. SE can be decreased by increasing the sample size, or decreasing the standard deviation.

The 95 % CI can be calculated for all sample statistics, but the calculations are relatively easy when the sample size is large such that the CLT can be invoked. For instance, suppose one is interested in calculating the 95 % CI for a true proportion. If the observed proportion is $p = r/n$, where r is the number of individuals in the

sample with the characteristic of interest in a sample of size n. In that case the 95 % CI is calculated as

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{n}},\ p + 1.96\sqrt{\frac{p(1-p)}{n}} \right].$$

In the above expression, $\sqrt{\frac{p(1-p)}{n}}$ is an estimate of the standard error of p under repeated sampling.

For example, in a study among 152 Brazilian pregnant women (Marin *et al*. (2005)), 71 (47 %) women had gingivitis. This is a relatively high percentage and one may wish to calculate the 95 % CI for the proportion of pregnant women with gingivitis in the population. The sample proportion is: $p = 71/152 = 0.47$. An estimated of the SE of the proportion is $\sqrt{\frac{0.47(1-0.47)}{152}} = 0.08$. Therefore, the 95 % CI $= 0.47 \pm 1.96 \times 0.08 = [0.31, 0.63]$. Accordingly, we can say that the true percentage of pregnant women with gingivitis in the population lies most likely between 31 % and 63 %.

10.7.6 Null- and alternative hypothesis

A (statistical) hypothesis is a statement of belief about population parameters. Hypothesis testing is a predominant feature of quantitative research in oral health and health care research in general. Given a sound theoretical structure, a representative sample from a population and the proper research design, researchers can test a hypothesis to see whether the collected data support or refute such hypothesis. There are two types of hypotheses: The null hypothesis, symbolized by H_0, proposes no relationship between two variables or no effect in the population. The alternative hypothesis, symbolized by H_A, is a statement that disagrees with the null hypothesis. If the null hypothesis is rejected as a result of sample evidence, then the alternative hypothesis is concluded. However, if the evidence is insufficient to reject the null hypothesis then it is retained, but not accepted. Traditionally researchers do not accept the null hypothesis from current evidence; they state that it cannot be rejected.

The steps of specifying a hypothesis will be discussed using data from a study by Warde *et al*. (2002) who conducted a randomized controlled trial (RCT) of oral pilocarpine in patients undergoing radiotherapy (RT) for head-and-neck cancer. The objective of this study was to test the hypothesis that the use of oral pilocarpine during and after RT for head-and-neck cancer would reduce the symptoms of post-RT xerostomia. Half of the subjects were randomly assigned to the treatment arm and the other half to the placebo arm. The primary outcome was the severity of xerostomia assessed by patients three months after RT, using 100 mm linear analogue scales. This scale assessed the patient's perception of dryness of their

mouth during the previous 3 days, with lower scores indicating the most difficulty. In this study the null and alternative hypotheses could be stated as follows:

H_0**:** The (population) mean of the ratings of dryness of the mouth among head and neck cancer patients is the same for patients who receive oral pilocarpine (μ_1) as those who do not receive oral pilocarpine (μ_2). Thus, $\mu_1 = \mu_2$.

H_A**:** The (population) mean of the ratings of dryness of the mouth among head and neck cancer patients is different for patients who receive oral pilocarpine (μ_1) from those who do not receive oral pilocarpine(μ_2). Thus, $\mu_1 \neq \mu_2$.

Denote $\mu_1 - \mu_2$ as Δ, then the above null- and alternative hypothesis can be reformulated as H_0: $\Delta = 0$ and H_A: $\Delta \neq 0$. In this example the direction for the mean difference in ratings of dryness of the mouth (Δ) was not specified. In other words, it was not stated in the alternative hypothesis whether the mean ratings of dryness of the mouth among patients who receive oral pilocarpine will be higher or lower than that of patients who did not receive oral pilocarpine. This is known as a two-sided (or two-tailed) alternative hypothesis, which allows for the difference to be in either direction. Less often do researchers specify a one-sided alternative hypothesis in which a direction of effect is specified in the alternative hypothesis. An example of a one-sided alternative hypothesis is H_A: $\Delta < 0$. Using the above example this alternative hypothesis reads as: H_A: Head and neck cancer patients' mean ratings of dryness of the mouth are on average higher among those who receive oral pilocarpine (μ_1) than patients who do not receive oral pilocarpine (μ_2).

10.7.7 Hypothesis testing

The procedure of formulating a null- and alternative hypothesis was first suggested by Neyman and Pearson around 1930. They also proposed a decision rule to choose between H_0 and H_A. Their procedure has become the standard inferential statistical approach for choosing between two hypotheses. In the remainder of this chapter several examples of this procedure will be reviewed, we start with the comparison of two groups.

To test the null-hypothesis one employs a test statistic. The mechanism of testing the null-hypothesis will be illustrated with the study of Warde *et al.* (2002). Similar to the result that $\bar{X}$ has a normal distribution with mean μ and standard deviation $s/\sqrt{n}$, one can show that for large n_1 and n_2, $\bar{X}_1 - \bar{X}_2 = D$ has a normal distribution with mean $\mu_1 - \mu_2 = \Delta$ and standard deviation $s_D = \sqrt{s_1^2/n_1 + s_2^2/n_2}$. In the previous expressions the subindex refers to the treatment group. If the null-hypothesis is true then $\Delta = 0$ and this means that D must vary under repeated sampling around 0 with standard deviation s_D according to a normal distribution. Hence, if H_0 is true and under repeated sampling the majority of studies will

yield a value for D close to zero. In probabilistic terms: 95 % of the studies will give a value of D between $-1.96\,s_D$ and $1.96\,s_D$, or equivalently the standardized $Z = D/s_D$ lies with 0.95 probability between -1.96 and 1.96 if H_0 is true. For only 5 % of the studies Z lies beyond 1.96 in absolute value if H_0 is true. Now, if the observed standardized result lies beyond 1.96 in absolute value the researcher might decide that this result is too extreme to believe that H_0 is true. It is then said that H_0 is rejected at significance level $\alpha = 0.05$ or shorter at the 5 % level. When Z does not exceed 1.96 in absolute value, then accordingly, we do not reject the null hypothesis, and we say that the results are not significant at the 5 % level. This does not mean that the null hypothesis is true; simply that we do not have enough evidence to reject it.

Some clinical situations may require stronger evidence before rejecting the null hypothesis; researchers may decide to take $\alpha = 0.01$ or $\alpha = 0.001$. The chosen value for α must be taken prior to obtaining the results.

Observe that in practice the truth is never known and that therefore errors will be made by this decision rule (see also Section 10.7.9).

To illustrate the above concepts, the 2-IO Study is used. Apart from the baseline ratings, at two months post-treatment the participants were asked to rate again their level of satisfaction with their new prostheses using the same scale. Accordingly, the null and alternative hypotheses were stated as follows: H_0: The mean rating of general satisfaction among edentulous patients is the same in the two treatment groups and H_A: The mean rating of general satisfaction among edentulous patients is different in the two treatment groups.

In this study the sample size for the implant group (conventional group), n_1 (n_2), is equal to 54 (48) with post-treatment mean satisfaction rating equal to $\bar{x}_1 = 89.19$ ($\bar{x}_2 = 63.75$) and $s_1 = 20.3$ ($s_2 = 34.69$). This yields $z = d/s_d = 4.45$. Since the standardized difference lies beyond 1.96, the result is statistically significant at 0.05. Hence we reject the null-hypothesis and state that the mean satisfaction rating of the implant group is greater than that of the conventional group.

A variety of classical tests will be reviewed in the next sections and summarized in Table 10.2. As will be seen, the tests depend on the type of data collected (continuous or categorical variable) and on the kind of question we are interested. At the end of this chapter we will indicate the general mechanism of constructing and evaluating statistical tests.

10.7.8 The P-value

The Neyman-Pearson procedure describes, prior to collecting data, how the choice between two hypotheses needs to be made. Fisher, (one of) the most famous statistician(s) of the last century, suggested ten years earlier a measure aiming to show the evidence of the data against the null-hypothesis. This measure, called the P-value, can only be determined after having collected the data. Both the P-value and the Neyman-Pearson procedure are nowadays used concurrently, see Chapter 18 for a critical remark on combining these two approaches.

Suppose we have a study as in Awad *et al*. (2003) with exactly the same results, but based on smaller sample sizes (to illustrate better graphically the different concepts), namely $n_1 = n_2 = 10$. In that case, we would have found that $d/s_d = 2.00$ which is again too extreme to believe that H_0 is true. A measure of extremeness is the probability that a more extreme result could have been obtained if H_0 was true. In Figure 10.6 we have plotted the standard normal distribution and positioned the observed value for z for the fictive study. The probability to obtain a larger result than 2.00 is equal to 0.023, this is called the one-tailed/sided P-value. When extremeness is defined in a symmetrical manner, i.e. we add the probability to have a more extreme result than $-z$ in the other direction, then we speak of a two-tailed P-value, here equal to 0.046. Most statistical packages provide the two-tailed P-value automatically. Unless specified otherwise all, the P-values reported in the remainder of this chapter are two-tailed.

It can be seen from Figure 10.6 that the result is statistically significant at the (two-tailed) significance level 0.05, if and only if the (two-tailed) P-value is smaller than 0.05. This explains the notation $P < 0.05$ in case of a statistical significant result (at 0.05). Some papers only report whether the P-value is smaller than 0.05, 0.01, 0.001, etc. and indicate this with *, **, ***, etc. respectively and with 'NS' in case of a nonsignificant result (at 0.05). Summarizing the results by 'nonsignificant' or 'significant' reflects the approach of Neyman and Pearson (not-rejecting

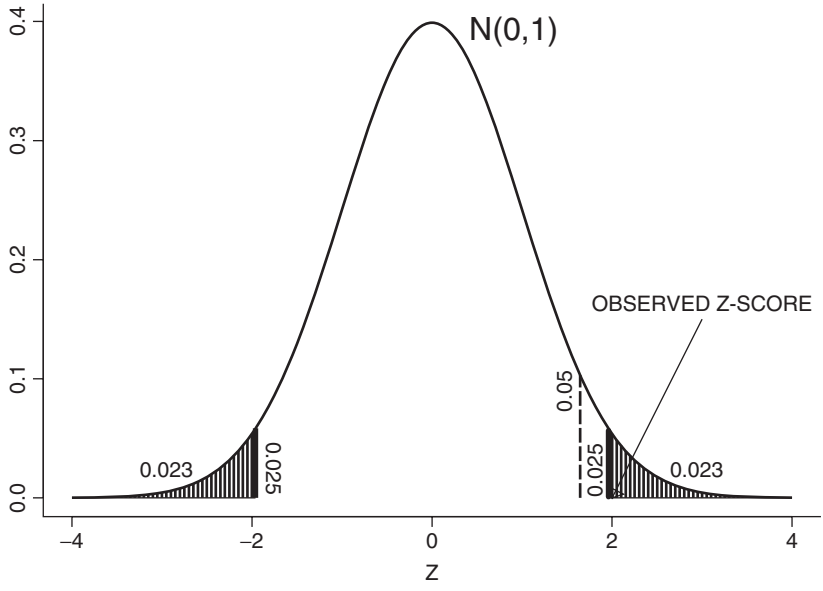


Figure 10.6 Standard normal distribution: 2-sided 0.05 rejection intervals (corresponding to vertical lines indicated with 0.025), 1-sided 0.05 rejection interval (corresponding to vertical line indicated with 0.05), grey areas represent two-sided P-value of observed result from fictive study.

or rejecting H_0). However, many researchers prefer to provide P-values in their reports to indicate the evidence of the data against H_0 which corresponds to the philosophy of Fisher.

Further, note an important relationship between the P-value and the 95 % CI which holds for many statistical tests. Namely, if the P-value for testing H_0: $\mu = \mu_0$ (or another parameter like proportion, etc.) is rejected at 0.05 then μ_0 is not contained in the 95 % CI for μ and vice versa.

Finally, the null hypothesis is either true or false. Hence, the P-value must not be interpreted as the probability that the null hypothesis is true, as is sometimes misleadingly claimed.

10.7.9 Errors in hypothesis testing

Most hypotheses tests compare groups of individuals who are exposed to a variety of experiments, as in Warde *et al*. (2002). On the basis of the observed results, two decisions need to be made: Reject the null hypothesis or do not reject the null hypothesis. In this way, two types of errors can be made in practice, since ... the truth is never known.

Type I error: A Type I error is committed when the researchers reject the null hypothesis when it is true. The chance for making a Type I error is equal to α. This means for the experiment of Warde *et al*. (2002) that the investigators would conclude that oral pilocarpine reduces the patients' perceptions of post RT xerostomia, when in reality it does not. Since in this study $\alpha = 0.05$, rejection of H_0 happens when $P < 0.05$. In fact, the obtained P-value obtained was greater than 0.05 and the authors did not reject H_0.

Type II error: When the null hypothesis is not rejected in the presence of a truly different effect of two treatments a Type II error is committed. In the experiment of Warde *et al*. (2002), it would mean that it is concluded that oral pilocarpine does not reduce patients' perceptions of post RT xerostomia when in reality it does. The chance of making a Type II error is denoted by β. Note that $(1 - \beta)$ is the power of the test, defined as the probability of rejecting the null hypothesis when it is false. Power is important in hypothesis testing and will be discussed in the next section.

10.7.10 Power of the test

At the planning stage of a study there should be a 'reasonable' chance of detecting a clinically relevant effect, if it exists. This a priori probability is, as seen above, the power of a test. In clinical trials classically the power is set to (at least) 80 %. Several factors affect the power of the study: (a) Sample size: the power increases with increasing sample size. A larger sample has a greater ability to detect a clinically important effect if it exists. When the sample size is small the test may have inadequate power to detect a particular effect. For example, in Warde *et al*. (2002) the authors estimated that 130 oral cancer patients (65 in each group)

are needed to obtain 80 % power to detect a reduction of 15 points in ratings of mouth dryness among patients randomized to the treatment group; (b) The effect of interest: the power of the test is greater for larger real effects; (c) The significance level: one is more likely to detect an effect with a significance level of 0.05 rather with 0.01.

10.8 Research questions with a numerical outcome

This section illustrates methods commonly used in oral health and medical research to analyze data obtained from studies of outcomes measured on numerical scales. First, two groups will be compared. In this case we will consider three cases: (a) the one sample t-test where an hypothesis is tested on one population mean; (b) the unpaired t-test, when the means of two independent groups are to be compared and (c) the paired t-test when means from two paired samples are to be compared. Secondly, we will compare the means of more than 2 groups.

10.8.1 The one sample t-test

When n is small we cannot rely on the CLT and we need to assume that the data have a normal distribution in order that $\bar{X}$ has a normal distribution under repeated sampling. However, even with a normal distribution for the data, the problem with a small sample size is that the standard deviation is estimated with greater error. This leads to replacing the standard normal distribution as reference distribution by the t-distribution. The t-distribution is symmetrical and varies with a parameter ν, called the degrees of freedom. Therefore, one speaks of a $t(\nu)$-distribution. For ν around 30 the t-distribution is close to the standard normal distribution (also called the Z-distribution). For smaller ν the tails of the $t(\nu)$-distribution become heavier. Also, the two-tailed 0.05 critical value of the standard normal distribution, 1.96 is replaced here by a value that depends on the degrees of freedom. For instance, for $\nu = 20$, 10, 5, 1 the two-tailed 0.05 critical value, denoted $t_{0.05}(\nu)$, of the $t(\nu)$-distribution is: 2.04, 2.09, 2.23, 2.57, 12.70, respectively.

The statistical test that arises from the above arguments is the t-test. The t test is used intensively in all areas of health research. We distinguish three types; here we consider the one-sample t-test. Although the t-test has been designed for small studies, in practice it is used for all sample sizes.

Mitchell *et al*. (2003) determined whether the dental caries activity among adolescents residing in Northern Manhattan, New York differs from that reported in the National Health and Nutrition Examination Survey III (NHANES III), to be considered the norm. Data were collected from 566 (thus a large sample size) 12 to 17 year old children at five school-based dental clinics in Northern Manhattan. The main objective of their study was to compare the mean DMFT (3.56) of children in their study to the reported mean DMFT (2.53) in the NHANES III study. Stated differently, they questioned whether the DMFT scores of children in their

study come from the same population as the NHANES III children. The hypotheses tested are now:

H_0**:** The mean DMFT of 12–17 years adolescents residing in Northern Manhattan equal to μ is not different from the norm, i.e. the mean in NHANES III equal to 2.53. Thus, H_0: $\mu = 2.53$.

H_A**:** The mean DMFT of 12–17 years adolescents residing in Northern Manhattan equal to μ is different from 2.53. Thus, H_A: $\mu \neq 2.53$.

Three factors play a role in deciding whether the population mean of the group under investigation differs from the norm: (a) the difference between the observed mean and the norm, (b) the amount of variability among subjects in the study and (c) the number of subjects in the study, n. The formula used to calculate the value of the test statistic is:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}, \tag{10.7}$$

and has a $t(n-1)$-distribution if H_0 is true. Here $t = (3.36 - 2.53)/(0.25/\sqrt{566}) = 79$. Referring to a $t(565)$-distribution, the P-value is much less than 0.01. Accordingly, we conclude that we have enough evidence to reject H_0 and hence we reject the hypothesis that the sample of DMFT values could come from a population with mean DMFT 2.53. Consequently, we infer that the mean DMFT for children in Northern Manhattan is different (in this case higher) from the mean DMFT in NHANES III.

One can also calculate the 95 % CI for the population mean μ, which is in general obtained from $\bar{x} \pm t_{0.05}(\nu) \times \left(s/\sqrt{n}\right)$. Since we are dealing here with a large sample size, $t_{0.05}(565) = 1.96$ and we obtain as 95 % CI for the population mean [3.34, 3.38]. Note that the 95 % CI does not include 2.53, confirming the relationship mentioned in Section 10.7.8.

For the t-test to be appropriate, observations should be normally distributed for n small. Histograms and box plots could be used to judge the shape of the data, but also more dedicated procedures to detect non-normality such as a normal probability plot and a normality test (e.g. Shapiro-Wilk test). Verifying normality might be important, since if the t-test is based on observations that are not normally distributed the P-value and the confidence interval might be seriously distorted. The good news is, however, that the t-test can still be used with data that deviate moderately from normality meaning that it allows to draw proper conclusions even when the distributional assumptions are not exactly met.

10.8.2 The unpaired t-test

If interest lies in comparing two different groups on some quantitative measurement, such as saliva flow, sensation of pain measured on a VAS scale, etc. we could use the t-test for independent groups, also called the unpaired t-test. The null- and alternative hypothesis are similar to those specified in Section 10.7.6 and the test

statistic T is similar to Z in Section 10.7.7, but the denominator is calculated somewhat differently. Further, T is referred to a $t(n-2)$-distribution with n the total sample size of the study (sum of sample sizes of the two groups).

The unpaired t-test is designed for small samples, but is applied in practice also to large sized studies. Two assumptions are involved with the unpaired t-test: (a) the data have in each group a normal distribution and (b) the variances of the two groups are equal. With equal sample sizes the latter assumption is less important, but with unequal sample sizes the reported P-value could be seriously distorted for unequal variances. In that case one uses a variation of the t-test allowing for unequal variances, where the degrees of freedom are not an integer anymore. Violation of the normality assumption can also affect the P-values and the confidence intervals. However, this issue is of less concern when sample sizes are larger than 30 in each group. When sample sizes are less than 30 in each group, then a nonparametric procedure called Wilcoxon rank sum test might be a better choice, see e.g. Dawson and Trapp (2004) or Hay (1997). Briefly, a nonparametric procedure replaces the observations by their ranks. For the Wilcoxon rank sum test the data are ranked globally and the group-specific ranks are compared. When there is severe imbalance in the two average ranks, the conclusion will be that the two distributions are different.

We repeated the analysis of Section 10.7.7 based on the data of the 2-IO Study and obtained $t = 4.58$, close to the $z = 4.45$ obtained above. The obtained t-value is referred to a $t(100)$-distribution and yields $P < 0.0001$. Again the null-hypothesis is rejected (at 0.05). The same is true for the t-test for unequal variances.

A 95 % CI for the difference in population means can be calculated using: $\bar{x}_1 - \bar{x}_2 \pm t_{0.05}(n-2) \times SED$, whereby SED is an expression of the standard error of the difference in means $(\bar{x}_1 - \bar{x}_2)$ different from that in Section 10.7.7 now using the assumption that the two variances are equal. In our example, the 95 % CI is [14.42, 36.46], not containing zero and thus confirming the significant result at 0.05.

10.8.3 The paired t-test

When measurements are taken for each individual on two different occasions, the scores are correlated and an independent t-test is not appropriate. For example, in the 2-IO Study, it could be of interest to compare pre-post treatment rating of satisfaction among participants in the study. In this case a paired t-test is the appropriate test when the differences have a normal distribution. It is easy to see that the paired t-test is in fact the one sample t-test with $\bar{X}$ replaced by the sample mean of differences pre-post and μ_0 equal to 0 if we wish to check that the two means are equal.

The null and alternative hypotheses are now: H_0: The mean difference in ratings of general satisfaction with mandibular prostheses before and after treatment in edentulous patients equals zero; and H_A: The mean difference in ratings of general satisfaction with mandibular prostheses before and after treatment in edentulous patients is different from zero.

The pre-treatment and post-treatment ratings of general satisfaction for all 102 patients were 38.49 (SD = 31.94) and 77.21 (SD = 30.64), respectively. However,

the paired t-test works on the differences. The average (pre-post) difference is -38.73 with SD $= 39.05$ and SE $= 3.87$. The T-value is equal to $-38.73/3.87 = -10.02$ and when referred to a $t(101)$-distribution a (highly) significant result at 0.05 is obtained ($P < 0.0001$). Accordingly, one can reject the null hypothesis and conclude that on average there was a significant improvement in ratings of general satisfaction with the new prostheses.

The paired t-test assumes that the difference (pre-post) has a normal distribution. If this is not the case then a non-parametric test can be applied, such as the Wilcoxon signed-ranks test.

10.8.4 Analysis of Variance (ANOVA)

When interest lies in the comparison of more than two groups, one might apply repeatedly the unpaired t-test comparing the k groups pairwise. This approach is, however, not appropriate as it leads to the problem of multiple testing, which is explained with the following example. Sanders *et al*. (2006) compared the number of missing teeth among Australian adults according to five categories of socioeconomic status (low, low to moderate, moderate, moderate to high and high). To find out whether there is a difference between some of the $k = 5$ groups in the average number of missing teeth, one could apply $4 \times 5 = 20$ unpaired t-tests comparing group 1 versus 2, 1 versus 3, etc. Now suppose that the null hypothesis is true, i.e. all 5 true means are equal. The above procedure implies that 20 unpaired t-tests are applied each time under H_0. Each time there is a risk of rejecting H_0 while it shouldn't be rejected. The risk of wrongly rejected H_0 with one t-test is 0.05, when the significance level $\alpha = 0.05$. The risk of wrongly rejecting H_0 with two t-tests is a bit less than 0.10. Thus, although the risks do not simply add up, the total risk (called the experimentwise error rate) of rejecting H_0 with 20 pairwise t-tests will be much greater than 0.05 (called the nominal error rate) and in that case one speaks of an inflated probability of Type I error.

One-Way Analysis of Variance (ANOVA) is the appropriate approach for analyzing data that compares means of $k > 2$ groups of independent observations. This method protects against the Type I error inflation by assessing if any differences exist at all among (true) group means. A significant ANOVA result only tells us that there is somewhere a difference between the group means, but we do not know yet where that difference is located. Dedicated pairwise tests can then be applied to find out where the significant difference is located. The topic of ANOVA is treated by many text books, see e.g. Dunn and Clark (1986), Berry *et al*. (2001) and Turner and Thayer (2001). In this section the basic concepts are discussed.

The principles and assumptions involved in One-Way ANOVA are similar to the unpaired t-test. For example, if three groups are compared we would like to know whether the observed differences between the three means are likely to have occurred by chance or something else played an important role (i.e. a treatment or an intervention). The assumptions involved in One-Way ANOVA are an extension of those of the unpaired t-test, i.e. (a) the outcome variable is assumed to have a normal distribution within each group and (b) the k population variances are equal.

Formally, with One-Way ANOVA one tests the following null- and alternative hypothesis:

H_0**:** The k population group means are equal, i.e. $\mu_1 = \mu_2 = \ldots = \mu_k$.

H_A**:** The k population group means are not all equal, i.e. there is at least one pair of groups (s,t) such $\mu_s \neq \mu_t$.

The term One-Way ANOVA has been chosen since the difference in the k means is tested by exploring variances. Indeed, in ANOVA the total variability in the data (over all groups) is split into two components: between-group and within-group variability. We speak of variability and not of variance, because the above result holds for Sums-of-Squares (SS) which are strongly related but not exactly equal to variances:

(a) *Between-group variability*: The variability of the means of the different groups around the grand mean (calculated without taking into account the group structure). The corresponding Sum-of-Squares is denoted SS_B.

(b) *Within-group variability*: This is also referred to as the unexplained or residual variation and refers to the random variation between individuals within each group. The corresponding Sum-of-Squares is denoted SS_W.

One can show that $SS_T = SS_B + SS_W$, with SS_T the total variability of the data around the grand mean. Accordingly, if the total variability is similar to the within group variability, then the means of the k groups are not much different and hence there is not much between-group variability. The relationship between the total sum-of-squares and the ordinary variance obtained around the grand mean is given by: $MS_T = SS_T/(n-1) = s^2$ and is called the Total Mean Sum-of-Squares. There is also a Between-Group and Within-Group Mean Sum-of-Squares, i.e. $MS_B = SS_B/(k-1)$ and $MS_W = SS_W/(n-k)$, respectively. If H_0 is true then MS_B will be approximately equal to MS_W, while under H_A MS_B will be most often (much) greater than MS_W. Thus, an hypothesis test to test H_0 needs the ratio:

$$F = \frac{MS_B}{MS_W}.$$

Under H_0, this ratio has a F-distribution with $(k-1)$ and $(n-k)$ degrees of freedom denoted as the $F(k-1, n-k)$-distribution. To test H_0 at 0.05 significance level, one calculates F and compares it with the 0.05 upper critical value of the $F(k-1, n-k)$-distribution. If F exceeds this critical value, then one claims a significant difference at 0.05 among the k group means, otherwise the null hypothesis is not rejected. Alternatively, one calculates the P-value in a one-tailed manner, i.e. the area under the $F(k-1, n-k)$-distribution beyond the observed F-value to the right. If $P < 0.05$, then the result is called significant at 0.05.

In the 2-IO Study, data were collected at baseline on patients' preferred treatment. Patients were asked to indicate whether they had no treatment preference (neutral), preferred implants supported prostheses or conventional dentures. The

Table 10.1 2-IO Study: typical One-Way ANOVA table produced by SPSS on three group comparison

	Sum of squares	Df	Mean square	F	Sig.
Between groups	20611.93	2	10305.97	11.67	.000
Within groups	117471.69	133	883.25		
Total	138083.62	135			

researchers were interested in examining the association between patient preference and pretreatment ratings of general satisfaction. Three groups need to be compared: (1) 'neutral', (2) 'implant' and (3) 'conventional', hence $k = 3$. The group sizes are: $n_1 = 28$, $n_2 = 82$ and $n_3 = 26$, implying that $n - k = 133$.

In Table 10.1 a typical ANOVA table is shown. All components to calculate F are shown in the ANOVA table. A significant P-value equal to 0.00002 is obtained. Note that in the above ANOVA table a P-value of 0.000 is reported, which is impossible since a P-value can never be zero. Given the significant outcome, we conclude that at least two of the groups must have a different mean. The next question is then to locate these groups. Accordingly, Post-hoc (the Latin term means 'after this') comparisons are made or also called 'multiple comparison tests'. Most of these tests are designed to perform two-group comparisons in a valid manner, i.e. controlling the probability of Type I error to $\alpha = 0.05$. There are several procedures on the market, such as Scheffé's test, Tukey's HSD test, etc. Using Scheffé's test, it was found that group 3 (participants who preferred conventional dentures) has a significantly higher average rating of pretreatment general satisfaction compared to the other two groups.

The ANOVA analysis is referred to as a One-Way ANOVA because the group means are lined up and compared. In Two-Way ANOVA there are two dimensions in the design of the study or in the analysis of the data. Suppose that in the 2-IO Study we wish to compare the three groups: 'neutral', 'implant' and 'conventional' but also wish to evaluate the impact of age class, say below or above 50 years of age. In that case, there are two factors playing a role. Factor 1 describes the patient's preference with 3 classes and factor 2 pertains to the age class of the patient with 2 classes. We may now wish to see whether there is an effect on the outcome due to factor 1 and/or due to factor 2. But one can be more ambitious here and ask whether the joint effect of the factors is simply the sum of the effect of each factor separately or in other words whether there is no interaction. There are now three F-tests to evaluate: (1) F-test to evaluate main effect of factor 1; (2) F-test to evaluate main effect of factor 2 and (3) F-test to evaluate interaction effect of factors 1 and 2. Typically, the last test needs to be performed first, if significant it implies that the effect of factor 1 differs in the different classes of factor 2, and one cannot speak of a main effect of each factor. We refer to textbooks in statistics for more details and more examples of the Two-Way ANOVA.

10.8.5 Transformations

Most of the statistical tests described above assume that the data have a normal distribution, such tests are called parametric (relying on a distribution specified by parameters, in this case μ and σ^2). If normality does not hold, non-parametric tests are an option (e.g. Wilcoxon rank-sum test, Wilcoxon signed ranks test, etc. as seen above) or transforming the data such that normality, and if necessary, other assumptions such as equality of variances are satisfied. When the original variable X is heavily right skewed then the distribution of $\log(X)$ will be more symmetric. Other transformations may be tried out, such as the square root transformation ($\sqrt{X}$), the inverse transformation ($1/X$), etc. However, in practice, it might not always be possible to transform the data such that normality is achieved for all groups, or that both equality of variances and normality are achieved. Further, working with transformed data might cause some difficulties in interpretation. A pragmatic approach, especially for small sample sizes might be to apply both a parametric and a corresponding nonparametric test. If both procedures result in the same conclusions, then the assumptions were not crucial, otherwise more work is needed.

10.9 Research questions with a nominal outcome

In the previous sections we were concerned with comparing group means. However, when the characteristic of interest is nominal, then the research question involves comparing proportions. We will consider here the comparison of two or more than two groups in this respect.

10.9.1 Testing a single proportion

In the study of Mitchell *et al*. (2003) the proportion of untreated dental caries among adolescents in Northern Manhattan (36 %) is compared to the national estimate of 16 %. This research question is translated in terms of statistical hypotheses as follows:

H_0**:** The true proportion of participants with untreated disease is equal to $\pi_0 = 0.16$.

H_A**:** The true proportion of participants with untreated disease is different from $\pi_0 = 0.16$.

To test H_0 we need a test statistic. For a large study, i.e. when n is large (rule of thumb is that $n \times p > 5$) we can invoke again the CLT. This leads to the following test statistic that can be used for large sample sizes:

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}, \tag{10.8}$$

where p is the observed proportion and π_0 the hypothesized proportion in the population. Because of the CLT we can refer Z to a standard normal distribution to determine the significance and the P-value.

For the above hypothesis test, we obtain $z = (0.36 - 0.16)/\sqrt{0.16(1 - 0.16)/566} = 0.2/0.015 = 13.33$. Referring z to the standard normal distribution, we see that it is much greater than 1.96 ($P < 0.0001$). The conclusion is that the proportion of untreated disease among Northern Manhattan adolescents is significantly greater than the national estimate. Note that the expression of the standard error of p in equation (10.8) differs from that used in the 95 % CI for the proportion. In the latter expression one imputes the observed proportion for π, while in (10.8) the value for π under H_0 is used.

If the sample size had been small, the above test is replaced by the binomial test, see e.g. Berry *et al.* (2001).

10.9.2 Comparing two independent proportions

Suppose we are interested in comparing two independent groups on a certain characteristic (discrete outcome). For example, we may be interested in comparing treatment preference (implant vs. conventional denture) among males and females in the mandibular 2-implant supported prostheses clinical trial as shown in Table 10.2. We can express the differential behavior with three measures: (1) absolute risk reduction $AR = \pi_1 - \pi_2$; (2) relative risk $RR = \pi_1/\pi_2$ and (3) odds ratio $OR = [\pi_1/(1 - \pi_1)]/[\pi_2/(1 - \pi_2)] = \pi_1(1 - \pi_2)/\pi_2(1 - \pi_1)$, whereby π_1 and π_2 are the true proportions in the two groups. These measures are estimated by replacing the true proportions by their estimates, i.e. p_1 and p_2. For Table 10.2, we obtain as estimates (for preference of implant with males representing the first group) $ar = 39/48 - 43/60 = 0.96$, $rr = 1.13$ and $or = 1.71$, respectively. All three measures show that males have a higher preference for implants than females. The question is whether this preference is statistically significant.

Suppose that the study is large, whereby large will be defined below. In this case we can conduct a chi-square test to test the following hypotheses:

H_0: Preference for conventional dentures is the same in edentulous men (π_1) and women (π_2), i.e. $\pi_1 = \pi_2$.

H_A: Preference for conventional dentures is not the same between edentulous men and women, i.e. $\pi_1 \neq \pi_2$.

Note that H_0 can also be restated as 'factor gender is independent of the factor preference'. Hence in H_0 we actually claim that there is independence between gender and preference. In Table 10.2 we show the frequency of patients that participated in 2-IO Study, split up according to gender and preference of treatment. This table is called a 2×2-contingency table, since the basic table (without the extra calculated numeric values) is built up of two rows and two columns. The observed frequencies in the table are denoted O_{ij}, whereby i refers to the ith row and j to the jth column. Hence, $O_{12} = 17$ is the frequency at row 1 and column 2.

Table 10.2 2-IO Study: gender by preference (conventional versus implant) cross-tabulation

			Gender		
			Male	Female	Total
Preference	Conventional	Count	9 (O_{11})	17 (O_{12})	26
		Expected Count	11.6 (E_{11})	14.4 (E_{12})	26.0
	Implant	Count	39 (O_{21})	43 (O_{22})	82
		Expected Count	36.4 (E_{21})	45.6 (E_{22})	82.0
	Total	Count	48	60	108
		Expected Count	48.0	60.0	108.0

To test H_0 we calculate the frequencies we would get if H_0 were true, these are called the expected frequencies and denoted E_{ij}. The calculation of the expected frequency is quite simple:

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}.$$

As an example, $E_{11} = 26 \times 48/108 = 11.6$. The expected frequencies are most often not integers and should be interpreted as the expected numbers of patients in the respective cells of the 2×2-contingency table under repeated sampling if H_0 were true. If the expected and observed frequencies differ a lot, then this is an indication that H_0 is most likely not true. The test statistic is:

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{10.9}$$

which follows a χ^2-distribution with 1 degree of freedom under H_0. The $\chi^2(1)$-test for independence is a one-tailed test, the 0.05 critical value of the $\chi^2(1)$-distribution is 3.841. Therefore, we will reject the null hypothesis of independence if X^2 in (10.9) is greater than 3.841. For our example we obtained $X^2 = 1.34$ corresponding to $P = 0.25$ and hence we do not reject H_0.

The chi-square test is applied for large studies. Large now means that all expected cell frequencies are greater than 5, which is the Criterion of Cochran. When this criterion is not satisfied, we are not sure that the reported P-value is correctly calculated. There are two alternative procedures: (1) a χ^2-test with continuity correction and (2) Fisher's Exact test, see Berry *et al*. (2001) for more details on these tests.

Finally, one can also construct a 95 % confidence interval for the difference $\pi_1 - \pi_2$. Here we obtained as 95 % CI : $[-6.27, 25.47]$. Note, that the upper and lower limits of the 95 % CI overlap with 0 and hence we can decide also from this result that there is no significant association (at 0.05) between gender and preference for treatment.

An alternative testing procedure is based on the large sample (also called asymptotic) behavior of one of the above three measures (ar, rr, or) in case H_0 is true.

Take the test based on the odds ratio. Under H_0 the logarithm of or, $\log(or)$ has a normal distribution with mean 0 and variance $v^2 = 1/O_{11} + 1/O_{12} + 1/O_{21} + 1/O_{22}$. Then $z = or/v = 0.54/0.47$ corresponds to $P = 0.25$ (the same as with the $\chi^2(1)$-test) when referring to a N(0,1) distribution.

10.9.3 Comparing two dependent proportions

If interest lies in comparing the proportion of subjects with or without a characteristic after an intervention or with passage of time, then the McNemar test is recommended. This test is similar to the paired t-test used with numerical data. An example of the use of this test in the dental literature is illustrated in Schaeken *et al*. (1991). In this study, the authors compared the color of root surface lesions among 44 patients with advanced periodontal disease at baseline (total light color 68 % and total dark color 32 %) and one year post-treatment (total light color 80 % and total dark color 20 %). Now the null hypothesis is that the proportion of lesions with different colors is the same before and after treatment. The alternative hypothesis is that these paired proportions are not the same. The test statistic that pertains to the McNemar test is different from that in (10.9) but uses also the $\chi^2(1)$-distribution as reference. The results from this study for the McNemar test showed that the P-value was equal to 0.025, indicating that the McNemar test value was greater than the critical value of 3.84. Therefore, the authors rejected the null hypothesis and concluded that there is a difference in lesion pairs.

10.9.4 Comparing independent proportions in more than two groups

Consider the example of Section 10.9.2 taken from the 2-IO Study, but with preference extended to a larger number of categories: (1) no preference, (2) preference for conventional dentures and (3) preference for implant overdentures. The following null and alternative hypotheses apply:

H_0**:** Treatment preferences are independent from gender.

H_A**:** Treatment preferences are not independent from gender.

The test statistic for testing H_0 is similar to that in (10.9), i.e.:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{10.10}$$

where r is the number of rows and c is the number of columns. Now X^2 is referred to a $\chi^2((r-1)(c-1))$-distribution, whereby $(r-1)(c-1)$ are the degrees of freedom, i.e. df. For this example this means that $df = 2$. The 0.05 critical value of the $\chi^2(2)$-distribution is 9.21. Since the calculated $X^2 = 3.685$, we cannot reject H_0 and one has no evidence of an association between preference and gender.

Table 10.3 Overview of statistical tests comparing 2 or more than 2 groups for numerical and nominal outcomes. The tests indicated with an asterisk can also be used for ordinal outcomes ("corr" stands for "corrected")

Comparison of groups				
k			Sample size	
	Type comparison	Distribution	Large	Small
	Unpaired comparison (independent groups)			
	Means	Normal/Var =	Unpaired t-test	Unpaired t-test
	Means	Normal/Var $\neq$	t-test for $\neq$ var	t-test for $\neq$ var
	Means/ distributions	Not-normal	Unpaired t-test	Wilcoxon-Mann-Whitney test*
2	Proportions		χ^2-test	Corr χ^2-test/ Fisher's Exact test
	Paired comparison (dependent groups)			
	Means	Normal	Paired t-test	Paired t-test
	Means/ distributions	Not-normal	Paired t-test	Wilcoxon signed-ranks test*
	Proportions		McNemar test	Corr McNemar test/ Binomial test
	Independent groups			
	Means	Normal/Var =	1-Way ANOVA	1-Way ANOVA
	Means/ distributions	Not-normal	1-Way ANOVA	Kruskall-Wallis test*
> 2	Proportions		χ^2-test	Corr χ^2-test/Exact test
	Dependent groups See next chapters			

10.10 Overview table statistical tests

In Table 10.3 the tests are tabulated according to (1) the number of groups compared; (2) the type of comparison (independent versus dependent); (3) whether numerical rather than categorical data are used; (4) the sample size of the study and (5) the assumptions and whether they are satisfied.

10.11 The likelihood

In the study of Mitchell *et al*. (2003) 36 % of the 566 adolescents in Northern Manhattan were found to have untreated dental caries. Assuming that the adolescents are independent of each other, the probability that $X = 204$ among $n = 566$ is described by expression (10.4) of the binomial distribution. For each value of π expression (10.4) gives the probability that $X = 204$ happens. For $\pi = 0.1$ this probability is equal to 4.53×10^{-62}, quite small but we need to remember that

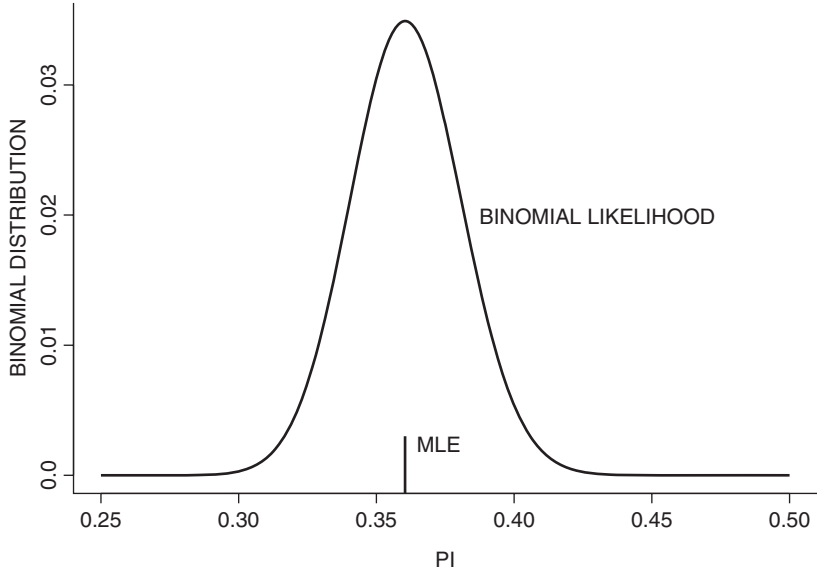


Figure 10.7 Study Mitchell et al. (2003): binomial likelihood function for probability of untreated dental caries among adolescents in Northern Manhattan.

the probability that any particular value $X = k$ happens will be very small since there are 567 possible values and $\sum_{k=0}^{566} P(X = k) = 1$. For $\pi = 0.3$ this probability becomes 3.01×10^{-4}, hence much greater. The largest value is obtained for $\pi = 204/566 = 0.36$. In that case the probability that $X = 206$ happens is equal to 3.49×10^{-2}. The probabilities $P(X = k)$ as a function of π is shown in Figure 10.7. This function is called the (binomial) likelihood function and expresses the probability of $X = 204/566$ for varying values of π. The function shows which values of π are plausible, given the obtained study results, and which are not. For instance, $\pi = 0.2$ seems to be very unlikely in the light of the data here. The function also shows the most likely value for π, the maximum likelihood estimate (MLE) of π, i.e. the value that corresponds to the highest likelihood value equal to $\hat{\pi} = 0.36$ here (the MLE is an estimator of π and therefore given a 'hat').

The likelihood concept and the maximum likelihood estimate are fundamental in statistics and many new statistical developments are obtained by writing down the likelihood of a model and exploring its characteristics. One of the most important results in statistics is that the MLE has asymptotically (for large n) a normal distribution with the correct value as mean and a variance that can be determined from the likelihood function. The general property that an estimator has the correct mean for large n is called consistency. In the above binomial likelihood this boils down to Z in expression (10.8) having a standard normal distribution, hence confirming what we already knew.

The asymptotic property of the MLE implies that the reference distribution for the (standardized) estimated regression coefficients in, say, logistic or Cox

regression (see Chapter 11), is the standard normal distribution under H_0. More specifically, with β such a regression parameter, $\hat{\beta}$ its MLE and var($\hat{\beta}$) its variance, then an often used statistic to test whether $\beta = 0$ is the Wald-statistic

$$\frac{\hat{\beta}}{\sqrt{\text{var}(\hat{\beta})}}, \tag{10.11}$$

which is referred to a N(0,1) distribution. For large n the Wald-statistic has the appropriate characteristics that a significance test should have, i.e. P(Type I error) $= 0.05$ (if $\alpha = 0.05$ is chosen). However, for small n P(Type I error) might not have the desired value. This is most often discovered from computer simulation studies whereby a computer program artificially generates studies (under H_0 and under H_A) and the empirical value of this probability, called the (actual) rejection rate, is determined. Similarly for the 95 % CI based on the MLE of β, i.e. $[\hat{\beta} - 1.96\sqrt{\text{var}(\hat{\beta})}, \hat{\beta} + 1.96\sqrt{\text{var}(\hat{\beta})}]$ we need that it covers the true value with probability 0.95. For small n the proportion of times the true value of β lies in the 95 % CI, the (actual) coverage probability, could be different however. Apart from the Wald-statistic, there is also the score-statistic and the likelihood ratio statistic. All three statistics aim to test model parameters, or the appropriateness of the whole model and will be used extensively in the subsequent chapters, even when not mentioned explicitly.

10.12 Miscellaneous topics

10.12.1 Superiority, equivalence and non-inferiority tests

All tests seen above are called superiority tests because they aim to show that two or more groups have a different mean or proportion. If the goal is to show that the groups show equal performance, then the superiority tests are not useful. Indeed, while in practice a non-significant result is often interpreted that the two or more groups are equal in performance, this conclusion is wrong. A common reason for a non-significant result is that the power is too low to detect any difference.

In case equal performance is the aim of the study, one needs an equivalence test. For this test one needs to define clinical equivalence, i.e. a $\Delta_E (>0)$ is needed such that two treatments with a treatment effect that differs at most with Δ_E units are considered to be equivalent for the patient. In that case, H_0 states that the groups differ in absolute value more than the clinical difference Δ_E and H_A states that the true difference is less than Δ_E. In case one wishes to show that an experimental treatment is not much worse than the conventional treatment, one needs a non-inferiority test. For this test one needs to define a clinically relevant upper bound to what can be tolerated as worse performance of the experimental versus the conventional treatment, i.e. $\Delta_{NI} (>0)$. Then H_0 states that the experimental treatment is more than Δ_{NI} units worse than the conventional treatment while H_A states that the experimental treatment is at most Δ_{NI} units worse than the conventional treatment.

It is beyond the scope of this chapter to discuss the statistical procedures corresponding to equivalence- and non-inferiority tests, see e.g. Lesaffre (2008) for a nonmathematical treatment of this topic. Finally, note that non-inferiority tests are becoming quite popular in RCTs.

10.12.2 Misuses of statistical tests and P-values

The P-value obtained from the sample statistics is compared to a predetermined α level and a decision is made accordingly to reject or not to reject H_0. However, the importance of the P-value is often overestimated, as it does not give any indication of the size of the effect and of the clinical importance of a particular treatment or intervention. Testing a new treatment against standard care and reporting a small P-value suggests that the observed difference is unlikely to be due to chance. However, to recommend the new treatment other factors should also be considered such as cost and side effects.

The scale of measurement determines the choice of statistical tests. For example, the χ^2-test is easily understood but sometimes used when another test is more appropriate. Indeed, often numerical variables are categorized and a χ^2-test is then applied. This leads to a loss of information and such practice is therefore not recommended. Unless there are clinical justifications and evidence that the selected categories are appropriate, statisticians caution against this practice, see e.g. Chen *et al*. (2007) and Royston *et al*. (2006).

The practice of what is referred to as 'data mining' can also lead to the use of inappropriate statistical methods. Researchers retrospectively analyze a collection of data using several statistical tests until a P-value is obtained that suggests certain relationships. In this situation data were not collected to test a particular hypothesis, but rather to search for a significant relationship. This is again an example of the multiple testing problem and will lead to many spurious relationships; there are ample papers in the medical and oral health literature that show violations against the multiple testing problem, see also Cloft (2006).

10.12.3 Statistical software

The illustrations in this chapter were obtained from the statistical software package SPSS® version 15.0 (SPSS Inc, Chicago, IL). There are hundreds of statistical software packages available on the market, some of which are freely available but most of them are commercial. They differ in (a) cost, (b) complexity, (c) user-friendliness, (d) whether the package is command- or menu-driven or a combination of both, (e) the easiness with which graphics can be obtained (and their quality), (f) the administration of data, etc. For the oral health researcher, the packages SPSS®, STATISTICA®, STATA® are a good choice since they are user-friendly and provide a rich set of possible procedures combined with easy-to-obtain graphics. Obviously, there are many more useful packages for the oral health researcher. Information from the above as well as other packages can easily be found by a Google search on the Web. The software *R* is freely available and is becoming

quite popular among statisticians. The package is command-driven, which means that there are no push buttons (in the standard version) to deliver statistical analyzes. However, there are numerous procedures written by users that are freely available and providing up-to-date statistical methods. For the more sophisticated oral health researcher this is definitely a package to explore. Finally, for the professional statistician there is the package SAS®, which is not only a statistical software package but also a database warehouse. Most likely this software is too demanding for most oral health researchers.

10.12.4 Complex structure of oral health data

The tests that we have reviewed in this chapter comprise simple comparisons between two or more groups. Regression models will be introduced in the next chapter. Dental data are, however, often much more complex in nature and need more sophisticated analyzes. In the next chapters techniques are discussed that take into account the hierarchical structure of dental data (mouth, teeth, surfaces), allow for dental data to be measured repeatedly, measured with error (misclassification and measurement error issues, interval-censoring), etc. We do not claim that such problems do not occur in other areas, but that it is only in oral health research that they come together!

References

Awad M, Lund J, Dufresne E & Feine J (2003) Comparing the efficacy of mandibular implant retained overdentures and conventional dentures among middle-aged edentulous patients: satisfaction and functional assessment. *Int J Prosthodont* **16**(2), 117–22.

Berry G, Mathews J & Armitage P (2001) *Statistical Methods in Medical Research (4th edition)*. Blackwell Scientific.

Chen H, Cohen P & Chen S (2007) Biased odds ratios from dichotomization of age. *Stat Med* **26**(18), 3487–97.

Cloft H (2006) The value of the P-value. *American J Neuroradiol* **27**, 1389–90.

Dawson B & Trapp R (2004) *Basic & Clinical Biostatistics (4th edition)*. Lange medical books, McGraw Hill.

Dunn O & Clark V (1987) *Applied Statistics: Analysis of Variance and Regression (2nd edition)*. John Wiley & Sons, Inc., New York.

Hay W (1997) *Statistics for the Social Sciences (5th edition)*. Holt, Rinehart & Winston.

Lesaffre E (2008) Superiority, equivalence and non-inferiority trials. *Bulletin of the NYU Hospital for Joint Diseases* **66**(2), 150–4.

Marin C, Segura-Egea J & Matinez-Sahuquillo A (2005) Correlation between infant birth weight and mother's periodontal status. *J Clin Periodontol* **32**(3), 299–304.

Mitchell D, Ahluwalia K & Albert, D. *et al.* (2003) Dental caries experience in Northern Manhattan adolescents. *J Public Health Dent* **63**(3), 189–94.

Naidu R, Prevatt I & Simeon D (2006) The oral health and treatment needs of schoolchildren in trinidad and tobago: findings of a national survey. *Int J Paediatr Dent* **16**(6), 412–18.

Royston P, Altman D & Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* **25**(1), 127–41.

Sanders A, Spencer A & Slade G (2006) Evaluating the role of dental behaviour in oral health inequalities. *Community Dent Oral Epidemiol* **34**(1), 71–9.

Schaeken M, Keltjens H & Van Der Hoeven J (1991) Effect of fluoride and chlorhexidine on the microflora of dental root surfaces and progression of root surface caries. *J Dent Res* **70**(2), 150–53.

Susin C, Haas A, Opermann R & Albandar J (2006) Tooth loss in a young population from Brazil. *J Public Health Dent* **66**(2), 110–15.

Turner J & Thayer J (2001) *Introduction to Analysis of Variance (1st edition)*. Sage Publications.

Warde P, O'Sullivan B & Aslanidis, J. *et al.* (2002) A phase III placebo-controlled trial of oral pilocarpine in patients undergoing radiotherapy for head-and-neck cancer. *Int J Radiat Oncol Bio Phys* **54**(1), 9–13.

11

Statistical methods for studying associations between variables

Brian G. Leroux

11.1 Introduction

Much of statistical methodology is concerned with making inferences about associations between two or more variables. For example, the Pearson correlation coefficient is a measure of the strength of the linear association between two quantitative variables. Regression is a class of powerful techniques for quantifying an association between a response variable (or outcome) and one or more explanatory variables (also called predictors, regressors, covariates, or independent variables). This chapter covers the basic concepts behind correlation and regression and emphasizes the interpretation of correlation and regression coefficients and the impact of various types of departures from model assumptions. Additional background on these topics can be obtained from texts on linear regression (Seber, 1984; Draper & Smith, 1998) and generalized linear models (McCullagh & Nelder, 1989).

Regression analysis may be used for two distinct purposes: (1) to gain an understanding of relationships between variables, and (2) to develop a model for prediction. For the first purpose, random assignment of treatment conditions to experimental units justifies interpretations of relationships in terms of causation (see Chapter 6). In contrast, observational studies, which do not employ randomization, allow inferences only about associations rather than causal effects.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

As a historical note, the term 'regression' comes from the phenomenon of the **regression effect** (also known as **regression to the mean**, see Chapter 9), which was first described by Galton in the context of associations between the heights of fathers and their sons. Galton noted the tendency for the average height of a son to be closer to the population average height than the height of his father. This pattern is observed for any two variables that are positively correlated with correlation value less than 1. A consequence of regression to the mean is that patients with extreme values of a clinical measure at one time will, on average, have less extreme values at a different observation time. Recognizing the regression-to-the-mean effect is critical to proper interpretation of changes in clinical measures over time. For example, if patients self-select to participate in a study based in part on poor current disease state, then the researcher will observe an improvement in patients' conditions in the absence of any intervention effect. As a consequence, inferring treatment effects based on changes over time is extremely dangerous; proper control groups are essential to demonstrating treatment effects.

11.2 Descriptive statistics for exploring associations

Descriptive statistical tools are useful for performing exploratory data analysis as a complement to more formal inferential methods that allow statistical inferences to be made about populations or treatment effects. Exploratory analyses may be targeted at checking data quality, uncovering patterns in data, or providing graphical illustration of such patterns. Tools of exploratory analysis relevant to the study of associations between variables include the scatterplot and the scatterplot smoother.

11.2.1 The scatterplot

The **scatterplot** is a basic technique for graphically depicting the association between two quantitative variables. For example, Figure 11.1 displays the average number of decayed, filled, or missing permanent teeth per child (DMFT) in each of 21 communities, plotted on the vertical (y) axis, against the corresponding values of the concentration of fluoride (ppm) on the horizontal (x) axis. These data were derived from the early studies by H.T. Dean in the 1940s (Chilton, 1967) on the association between caries and flouride concentration in drinking water. As displayed in Figure 11.1, large values of fluoride concentration tend to go along with small values of DMFT, whereas low concentrations of fluoride are associated with very high DMFT values. We say there is a *negative association* between fluoride concentration (x) and DMFT (y). In contrast, a *positive* association exists when large values of x tend to go along with large values of y.

Scatterplot smoothers The nature of the relationship between x and y can be described by superimposing on the scatterplot a smooth curve which describes the typical value of y at each value of x (Figure 11.1(b)). Each point on the smooth

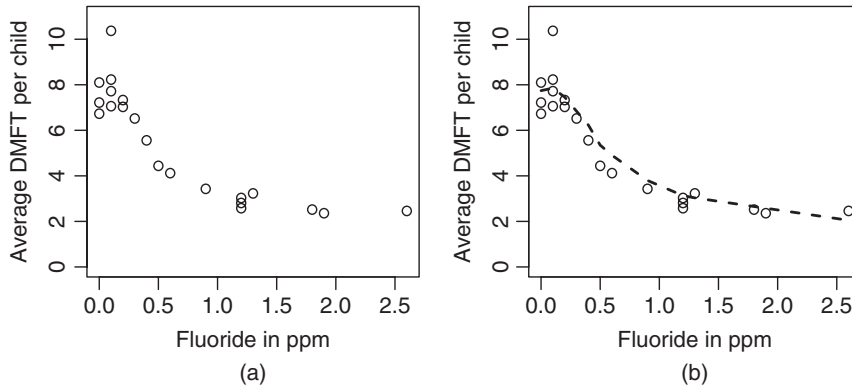


Figure 11.1 (a) Scatterplot of average number of decayed, filled or missing permanent teeth per child versus fluoride concentration (ppm) in the public water supply for 21 communities. (b) The scatterplot in (a) with the 'running lines' smoother (dashed line) superimposed.

curve in Figure 11.1(b) was obtained by forming a vertical window of values around a given position on the x-axis, fitting a straight line to points within this window, and then plotting a point at the given x-value and the fitted y-value determined by the line. The degree of smoothness of the curve is determined by the width of the vertical windows used (in this case the windows contained 40 % of the data points). The method used here is called the 'running-lines smoother' or 'super-smoother'. A smoother is useful for illustrating and identifying underlying relationships that cannot be described by a simple mathematical formula (such as a straight line or quadratic function). Linear regression, in contrast, imposes the constraint of a mathematical formula, but has advantages over smoothing in terms of interpretability.

Strength of the association A scatterplot is useful not only for highlighting a data trend, but also for displaying information about the strength of the association, i.e. to what degree the value of one variable can be predicted from the value of the other. For the fluoride data in Figure 11.1, the association is quite strong, because the spread of the y values for a given value or narrow range of values of x is quite small compared to the overall spread of the y values. Another way of saying this is that the knowledge of the value of x for an individual community helps substantially in the prediction of the value of y, and that there remains only a small amount of variation around the predicted value. If the amount of variation around the smooth line is substantial compared to the amount of variation in y overall, we say there is a moderate association. A weak association is characterized by a minimal reduction in error. Note that the strength of association may not be displayed effectively in a scatterplot if the distributions of the variables are highly skewed or if inappropriate variable scales are used for the axes.

11.2.2 Correlation

The scatterplot and smoother are useful for illustrating the association between two variables and also providing some information about the strength of the association. It is sometimes helpful to have a quantitative measure of the strength of the association. One such measure is **Pearson's correlation coefficient**, denoted by ρ, which is a number between -1 and 1, with the value 1 representing perfect *linear* association, in which all the data points fall on a straight line having a positive slope. On the other hand, a correlation of -1 indicates that all the points fall on a straight line having a negative slope. Correlation values that are less than one in magnitude, either positive or negative, have similar interpretations except that the points do not adhere exactly to a straight line but are allowed to vary somewhat around a hypothetical straight-line relationship (Figure 11.2). The amount of variation around the line is reflected in the size of the correlation coefficient. Note that the value of the correlation coefficient does not depend on the units of measurement of the variables.

Interpretation of zero correlation The value of 0 for the Pearson correlation has a very special interpretation and is involved in one of the common fallacies in the use of correlation. The definition of zero correlation is no linear association, which means that the data points do not have any tendency to follow a *straight line* pattern with a non-zero slope. One example of the lack of a linear association is the pattern observed when plotting two variables that are **statistically independent**. For example, Figure 11.3(a) illustrates the scatterplot of 100 values of two independent random variables. The independence of the two variables is reflected in the scatterplot by the fact that the distribution of the values of y appears to be about the same for every position along the x-axis. This example illustrates the fact that two variables that are statistically independent must have zero correlation. The fallacy comes in assuming that the converse implication is also true, namely, that zero correlation implies statistical independence. In fact, zero correlation is consistent with many data patterns that are characterized by even very strong relationships, such as a quadratic relationship (Figure 11.3(b)). The fallacy associated

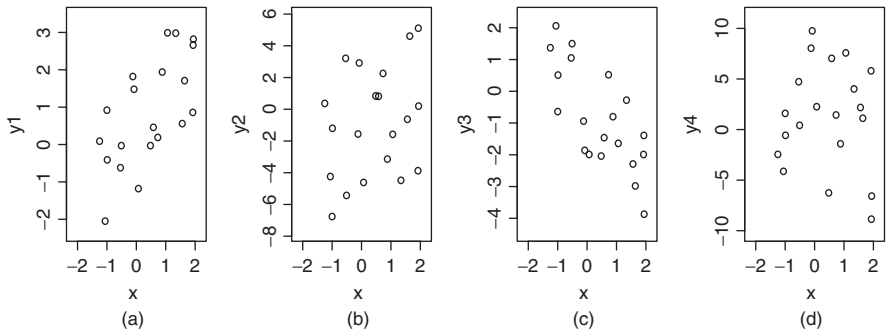


Figure 11.2 Scatterplots illustrating various values of the Pearson correlation coefficient: (a) 0.65, (b) 0.30, (c) -0.73*, (d)* -0.08*.*

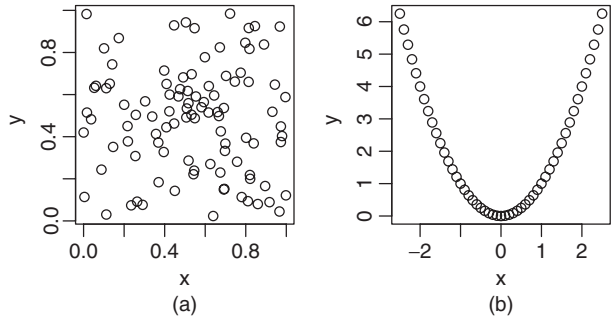


Figure 11.3 Illustrations of data patterns consistent with a Pearson correlation coefficient value of 0: (a) two independent uniformly distributed random variables; and (b) a quadratic relationship.

with zero correlation is one of the reasons it is so important to examine the data using scatterplots as a complement to correlation and regression analyses.

Calculation of the *Pearson correlation coefficient* The Pearson correlation coefficient is calculated using the following formula:

$$r = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}},$$

where $\sum_i$ denotes summation over $i = 1, \ldots, n$, as in $\sum_i x_i y_i = x_1 y_1 + \cdots + x_n y_n$, and $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively. An intuitive understanding of this expression is gained by considering the case when the variables have sample means ($\bar{x}$ and $\bar{y}$) equal to 0, and the sums of squared deviations from the means are equal to one, (i.e. $\sum_i (x_i - \bar{x})^2 = \sum_i (y_i - \bar{y})^2 = 1$. In that case, the expression for the correlation coefficient reduces to the sum of products $\sum_i x_i y_i$. From this expression, we see that observations with values of x_i and y_i that are either both positive or both negative will contribute a positive amount to the correlation, whereas a pair with opposite signs (one positive and one negative) will contribute a negative amount. Thus, the correlation reflects in part the relative frequencies of observations with the same versus opposite signs. The same holds in general in terms of the signs of the deviations from the mean. Given the simple mathematical expression ($\sum x_i y_i$) for r, it is hardly surprising that a given value, such as 0, could have alternative interpretations, as seen in Figure 11.3.

Variables with range restrictions The Pearson correlation coefficient is most useful for quantitative variables without range restrictions. Variables with severe range restrictions can yield correlation coefficients that are difficult to interpret. For instance, the interpretation of the correlation coefficient between two binary variables can be difficult because a binary variable can take on only two values (usually coded as 0 and 1). A numerical example will help to make this clear. Consider two binary indicators, representing presence or absence of two diseases

in an individual, and suppose that each disease occurs in 10 % of patients and that 3 % of patients have both diseases. Then the correlation between the two variables is 0.22, which suggests a weak association (compare with Figure 11.2). However, the odds-ratio (Chapter 10) between these two variables is approximately 5, that is, the odds for disease A are 5 times as high if the patient has disease B than if the patient does not have disease B. The fact that a strong association (as indicated by the odds-ratio of 5) is consistent with a small correlation of 0.22 indicates that one should be cautious when interpreting correlations for binary variables. Similar problems can occur for other types of variables with severe restrictions, such as highly skewed distributions that have many small values close to 0 as well as very large values that occur infrequently (e.g. person-level DMFT data).

Testing the significance of a correlation coefficient In applications of correlation it is often of interest to test the hypothesis that two variables are uncorrelated. This hypothesis is defined in terms of ρ, the population correlation coefficient between the two variables, namely the value of r obtained using all subjects in the population. A test of the hypothesis $H_0 : \rho = 0$ is performed by referring the test statistic

$$T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

to the t distribution with $n - 2$ degrees of freedom.

11.3 Simple linear regression

In modern statistics, regression refers to a very broad class of methods for describing the association between a set of explanatory variables and a single response variable. The basic feature of regression is a model that describes how the mean value of a response variable depends on one or more explanatory variables. For example, the equation

$$E[y|x] = \alpha + \beta x$$

postulates a straight-line relationship between the expected (mean) value of the variable y for a given value of variable x. For instance, if y represents the number of teeth lost by a dental patient with a given annual income x, then the above equation implies that the expected number of lost teeth per person, $E[y|x]$, is a linear function of income, x. The notation $E[y|x]$ represents the *conditional* expected value of the random variable y, conditional on the given value of x. The quantities α and β are unknown **parameters** of the regression model that would typically be estimated using data on a sample of individuals from the population.

To account for the variation in the response variable about the mean, the regression model is augmented by a random variable ϵ as follows:

$$y = \alpha + \beta x + \epsilon, \tag{11.1}$$

where y is the value of the response (number of lost teeth) for an individual patient, x is the patient's income, and ϵ is the amount by which the response deviates

from the average value for persons with income x. The deviation ϵ represents the combined influences on the number of lost teeth of all factors other than income; it is assumed to be statistically independent of income. The terms '**error**', '**noise**', and '**random variation**' are often used to describe ϵ, and its variance, σ^2, is referred to as the 'error variance'.

11.3.1 Least-squares estimates

The most commonly used method for fitting regression models to data is the **least-squares** (**LS**) method. The method is illustrated in Figure 11.4 using data on tooth loss and income for a sample of 80 patients. The fitted regression line has the equation $y = 3.5 + 0.043x$, where y represents tooth loss and x represents income in thousands of dollars.

The LS estimate of the slope β is calculated using the following formula:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{11.2}$$

where $\bar{x}$ and $\bar{y}$ are the sample mean values of the predictor and response, respectively. We use the 'hat'-notation ($\hat{}$) to denote an estimator of a parameter. After $\hat{\beta}$ is determined, the LS estimate of the intercept is calculated as follows

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \tag{11.3}$$

The defining property of the LS estimates is that the sum of squares of the residuals, $\sum_i e_i^2$, where $e_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$, is minimized. In Figure 11.4, the residuals are represented by the vertical distances from the individual data points to the regression line.

One of the appealing properties of LS is that the estimators $\hat{\alpha}$ and $\hat{\beta}$ are **unbiased**, which means that the average values of the estimates from an infinite set

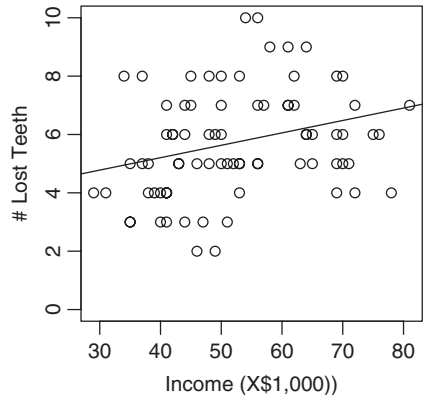


Figure 11.4 Scatterplot of number of lost teeth versus income and LS regression line defined by the equation $y = 3.5 + 0.043x$.

of repeated samples from the population would be equal to the true values of the parameters. Unbiasedness is a desirable property for an estimator to possess; however, its importance is limited by the fact that it depends on the assumption that the model is true. Linear regression models, like statistical models in general, are usually not thought of as perfect reflections of reality but only as useful approximations. Therefore, it is essential to assess how regression analysis behaves in the presence of departures from the model assumptions. This will be done in Section 11.5.

11.3.2 Interpretation of the regression coefficients

The interpretation of the regression coefficient β in (11.1) is the *average difference between responses for two individuals per unit difference in their values of* x. Thus, in the tooth-loss example, the interpretation of the estimate of the regression coefficient $\hat{\beta} = 0.043$ is that the average difference in number of lost teeth for two individuals is 0.043 per \$1,000 of difference between their incomes. In other words, each additional \$1,000 of income is associated with an increase of 0.043 teeth lost. The interpretation of the estimated intercept $\hat{\alpha} = 3.5$ is that the average number of lost teeth for those with no income is estimated to be 3.5. It is important to recognize that the association between income and tooth-loss may not be causal, that is, it may be due to risk factors for caries and periodontal disease being particularly prevalent in those with high incomes, rather than a direct causal effect of income on tooth loss.

There is an implicit assumption made in the interpretation of the regression coefficient given above: in the comparison of two individuals, the average difference in responses depends only on the *difference* between values of the predictor (and not on their absolute values). This assumption may fail, for example, if the difference between number of teeth lost for two individuals with a difference of \$1,000 in incomes is smaller at the upper end of the income scale than at the lower end. In practice, we interpret β in an average sense, that is, *averaged over the population of predictor values*. If the regression model has been misspecified, the value of β in another population may be different simply because the range of values of predictors changes, even if the underlying relationship between response and predictor is the same in the two populations. Model misspecification creates even greater difficulties for multiple regression models, as will be seen in Section 11.5.

11.3.3 Statistical inference for the regression coefficient

In addition to obtaining an estimate of the regression coefficient, it is essential to provide an assessment of the uncertainty in the estimate. The usual uncertainty assessment is the standard error of $\hat{\beta}$, calculated using the formula

$$\text{SE}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

This formula depends on an estimate of the **error variance**, which is given by $\hat{\sigma}^2 = \sum_{i=1}^{n} e_i^2/(n-2)$, where $e_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$ is the ith **residual**. In words,

the estimate of the error variance is roughly the average of the squares of the deviations of the responses (y_i) from their corresponding **'fitted' values** ($\hat{\alpha} + \hat{\beta}x_i$).

Using the standard error, a **confidence interval** for β is obtained as $(\hat{\beta} - t_{n-2,1-\alpha/2}\text{SE}(\hat{\beta}), \hat{\beta} + t_{n-2,1-\alpha/2}\text{SE}(\hat{\beta})$. A test of the null hypothesis of no association between x and y, i.e., $H_0 : \beta = 0$, can be performed by referring the t-statistic $T = \hat{\beta}/\text{SE}(\hat{\beta})$ to the t_{n-2} distribution (see Table 11.2 for examples). The test of significance of β is equivalent to the test of zero correlation given previously (note that the hypotheses $H_0 : \rho = 0$ and $H_0 : \beta = 0$ are equivalent).

11.3.4 Using the regression line for prediction

In applications of regression to prediction, interest lies in the distribution of the response y for a given value of the predictor x. In the income versus tooth loss example, a patient with income of \$30,000 would have a predicted number of lost teeth given by $3.5 + 0.043 \times 30 = 4.8$. In order to be useful, there must be an assessment of the likely prediction error (difference between the actual number and the predicted value). Statisticians typically use the variance of the prediction error for this purpose (just as a variance is used for quantifying the uncertainty in a parameter estimate). The following formula gives the variance of the prediction error for a simple linear regression model:

$$\text{Prediction Error Variance} = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

This expression accounts for both the uncertainty in the estimation of the parameters of the regression model, and the inherent variability in the value of y to be predicted. Note that the variance increases as the value of the predictor variable x moves away from the sample mean $\bar{x}$.

11.4 Multiple regression

Simple linear regression is a powerful method for describing the association between an outcome variable and a single predictor variable, but it cannot describe associations involving three or more variables together. This type of association arises when there is a need to adjust for a covariate that may be acting to confound an association of interest. For example, in an observational study of the association between an exposure and a measure of disease, it is often necessary to control demographic and other variables that may be confounders (Chapter 7). Associations involving three or more variables also arise in the development of models for predicting an outcome using more than one predictor variable. **Multiple regression** is one very useful method for these aims.

11.4.1 Multiple regression for controlling confounders

Recall the *simple* linear regression model given by Equation (11.1), which describes the average number of lost teeth as a function of annual income. Suppose that a

patient's age is related both to income and the number of teeth lost; in epidemiologic terminology, age is described as a confounder of the association between income and tooth-loss. In order to estimate an income-tooth loss association that is not confounded by age, but is instead *independent of* age, we add a term representing age to the model, which gives the following *multiple regression* model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where x_1 represents income (in 1,000s of \$) and x_2 represents age of the patient in years. The interpretation of β_1 in this model is quite different than the interpretation of the coefficient of the income variable in the simple linear regression model (11.1). In particular, β_1 is the average difference in teeth lost per unit difference in income, *for two individuals with equal ages*.

Table 11.1 shows the results of fitting the multiple regression model to the tooth loss data (Model II) and compares these results to those obtained previously by fitting the simple linear regression model without the age variable included. Note that the estimate of the coefficient for income in the multiple regression model has the opposite sign as the coefficient in the simple model. The simple linear regression model indicates that a person is expected to have more teeth lost than a different individual with a lower income. In contrast, the multiple regression model says that, if we compare two individuals *with equal ages*, then each additional \$1,000 of income is associated with an average *decrease* of 0.083 teeth lost. The coefficient -0.083 describes the association between tooth loss and income, *controlling age*. An alternative approach to controlling age is to perform an analysis on a subset of individuals that is relatively homogeneous on age. For example, performing simple linear regression of tooth-loss versus income for individuals aged 65 years and older gives an estimated coefficient for income equal to -0.092, which is quite different than the coefficient of 0.043 from the entire sample, but similar to the result for the multiple regression model.

The results for the tooth-loss data illustrate the dramatic changes that can occur in a regression coefficient when the model is modified, such as by the addition of another explanatory variable. Other modifications, such as applying a transformation (e.g. square-root or logarithm) to one of the variables, can also have dramatic results.

Table 11.1 Multiple regression results for the tooth-loss data. The results are presented using the following format: Estimate $\pm$ Standard Error (p-value).

Variable	Model I	Model II
(Intercept)	3.50 ± 0.89 (<0.001)	1.10 ± 1.10 (0.32)
Income	0.043 ± 0.017 (0.01)	-0.083 ± 0.041 (0.04)
Age	NA	0.158 ± 0.047 (0.001)

11.4.2 General form of the multiple regression model

One can have more than two explanatory variables in a regression model. The general form is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

where $x_1, x_2, \ldots, x_p$ are the values of p explanatory variables used to describe the mean of the response variable y. As with simple linear regression, the error term ϵ is assumed to have mean 0 and constant variance σ^2, and to be statistically independent of the explanatory variables. The interpretation of a coefficient in this model, such as β_1, is the average difference in responses per unit difference in x_1, with all other variables, $x_2, \ldots, x_p$, held fixed. As the tooth-loss data example shows, the value of a regression coefficient can be greatly affected by the presence of other variables in the model.

11.4.3 Multiple regression for prediction

Multiple regression is also useful for developing models for prediction. For example, in the context of the data on fluoride and caries shown in Figure 11.1, it may be useful to derive a model for prediction of average DMFT in a community based on fluoride concentration (F). As depicted in Figure 11.5(a) the relationship between DMFT and fluoride concentration is not described well by a straight line; therefore, a simple model of the form, $DMFT = \alpha + \beta F + \epsilon$ would not yield good predictions in many cases. One way to improve the model is to use a polynomial function of F as the predictor, for example, a second-order (quadratic) polynomial model would have the form $DMFT = \alpha + \beta_1 F + \beta_2 F^2 + \epsilon$. In this application of multiple regression, the different explanatory variables are simply different **transformations** of a single predictor variable F. Comparing

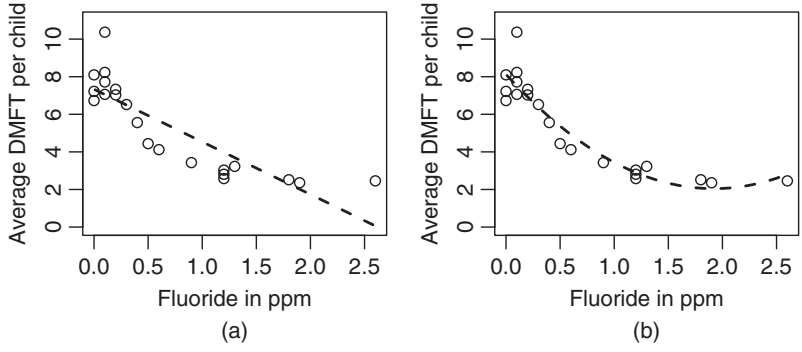


Figure 11.5 Fitted curves for the fluoride data based on (a) a model that is linear in fluoride concentration; and (b) a model that is quadratic in fluoride concentration.

Table 11.2 Results from fitting straight-line and quadratic models to the fluoride data.

	Straight-line model			Quadratic model		
Variable	Estimate	SE	p	Estimate	SE	p
(Intercept)	7.332	0.390	<0.0001	8.126	0.322	<0.0001
F	−2.797	0.386	<0.0001	−6.333	0.815	<0.0001
F^2				1.649	0.359	0.0002

Figures 11.5(a) and 11.5(b) shows that the quadratic model fits the data better than the straight line. The improvement in fit is reflected in the statistical significance of the quadratic term (Table 11.2), and can be quantified as a reduction in the **residual standard deviation** ($\hat{\sigma}$) from 1.29 for Model I to 0.90 for Model II, which is a reduction of 30 %. Further improvements in fit can be achieved either by adding additional powers of F to the right-hand side of the equation, or by a transformation (e.g. logarithm) of the response variable.

11.4.4 Multiple correlation

The **multiple correlation coefficient** is a measure of how well a response variable is predicted by a set of two or more predictor variables. Generalizing Pearson's correlation coefficient for a single predictor variable, the multiple correlation coefficient, denoted by R^2, is calculated from the following equation:

$$1 - R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\hat{y}_i$ is the predicted response for the ith observation, calculated as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}$. The interpretation of R^2 is similar to that of the Pearson correlation coefficient. For instance, an R^2 equal to 1 implies a perfect linear relationship between y and the set of predictor variables. The minimum value of R^2 is 0, which is consistent with y being statistically independent of the predictor variables, but is also consistent with many different non-linear relationships as was shown previously for the case of a single predictor variable (Figure 11.3(b)). In general, R^2 is interpreted as the fraction of the variance of y that is explained by the predictor variables. The values of R^2 for the linear and quadratic models fit to the fluoride data are 0.73 and 0.88, respectively.

11.4.5 Partial correlation

In some applications, it is of interest to quantify the strength of association between the response and one of the predictors, controlling the other predictors. This type of association is quantified by the **partial correlation**. To illustrate this concept

we use the tooth-loss data. First, consider the matrix of correlations for income, age, and tooth-loss:

	Income	Age	Tooth-loss
Income	1.00	0.92	0.28
Age	0.92	1.00	0.39
Tooth-loss	0.28	0.39	1.00

This shows that the correlation between tooth-loss and income is equal to 0.28, and that there are positive correlations between age and each of the other variables: 0.92 for income versus age and 0.39 for tooth-loss versus age. As we have already seen, the association between income and tooth-loss is dramatically changed by controlling age (the regression coefficient changed from a positive 0.043 to −0.083). Similarly, the correlation measure of the strength of association between income and tooth-loss is changed dramatically by control of age; specifically, the partial correlation between income and tooth-loss, controlling age, is equal to −0.22, which is calculated as follows using the elements of the correlation matrix given above:

$$\frac{0.28 - 0.92 \times 0.39}{\sqrt{1 - 0.92^2}\sqrt{1 - 0.39^2}} = -0.22.$$

This value indicates that controlling age changes the sign of the association between income and tooth-loss and also slightly weakens the strength of the association (the absolute value of the correlation decreased from 0.28 to 0.22). Partial correlations can also be obtained with control of more than one variable, although the calculation is more complicated than the simple formula shown above.

11.5 Model mispecification

Regression modelling is a very powerful tool for making statistical inferences, but one that must be used carefully to avoid invalid conclusions. This section will first consider the effects of the different types of model departures on the validity of regression analyses. Although this material is described in terms of linear regression, much of it is applicable also to the advanced regression models introduced in Section 11.6. After that, we present an introduction to the use of residual diagnostics for detecting the various model departures and an introduction to influential observations.

The following discussion focuses on the effects of model misspecification on inferences made about the regression coefficients. Different considerations apply in the context of using regression for making predictions of future responses. For instance, in that setting, the normality assumption becomes important even for large sample sizes.

11.5.1 Impacts of model misspecification

We consider the standard assumptions of linear regression in increasing order of the potential for invalid conclusions to result from an incorrect assumption: (1) normal distribution and constant variance, (2) independent observations, and (3) correctly specified mean model.

Normal distribution and constant variance. The assumption of a normal distribution for the errors is the least critical of the regression model assumptions. To be specific, inferences about regression coefficients from a regression analysis are valid for any error distribution provided only that the sample size is large enough. The theoretical justification of this statement is the **Central Limit Theorem** (Chapter 10). Fortunately, the sample size requirement for the validity of regression analysis is not very great in most practical situations (Lumley *et al*., 2002). More important than the assumption of normality is the assumption of a **constant variance** for the errors. Fortunately, the **robust variance estimate** (Chapter 13) yields valid inference even in the presence of non-constant variance.

Independent responses. A critical assumption of the multiple regression model is the statistical independence of the observations. Departures from this assumption are very common in oral health research, because multiple observations are often taken on each patient. For example, in periodontal research, clinical attachment levels may be recorded on a large number of sites in each patient. The application of regression analysis to such data, without proper accommodation for the **dependence** between responses for the same patient, can lead to very misleading results, including hypothesis tests about regression coefficients that are either overly conservative (p-values too large) or highly anti-conservative (p-values too small). Correlated responses present serious challenges to analysis of oral health data (Chapter 13).

Correctly specified model for the mean response. In most situations, the most critical assumption of multiple linear regression is the assumption that the form of the model for the mean response is correctly specified. This assumption is particularly important for applications of regression to the discovery of associations between exposures and a response variable that may represent underlying causal relationships. In this context, if the regression model is misspecified, such as by the omission of an important confounder variable, the estimated association between exposure and response is a biased estimate of the underlying causal relationship. Although the other three types of misspecification described above affect statistical inferences about the coefficients, misspecification of the model for the mean response is more serious because it causes the estimated coefficients to be biased (Chapter 10). Bias is such a serious concern because it represents error in estimation that persists even for very large sample sizes, unlike random error due to sampling variability which decreases with increasing sample size. Modern statistical methods have been developed to effectively mitigate the problems associated with nonnormality, nonconstant variance, and correlated responses; however, it may be impossible to correct a misspecified model if an important covariate has been omitted and is unknown or unmeasured.

The income-tooth loss example results of Table 11.1 provide an illustration of the dramatic effects of an omitted covariate. If Model II was correct but we fit Model I, which excludes the age variable, then we would conclude that income had a positive association with tooth loss. This conclusion would be very misleading if applied inappropriately, for example, assumed to imply a causal relationship, or applied to a new population that may not have the same distribution of age and income levels. Because of the potential for misleading results, the choices of explanatory variables to include in a regression model, and possible transformations, are critical. These choices must be made on scientific grounds and should not be based on purely statistical considerations such as those used in **stepwise regression**.[1] Finally, because an omitted covariate is always a possibility, regression results must be interpreted with extreme care.

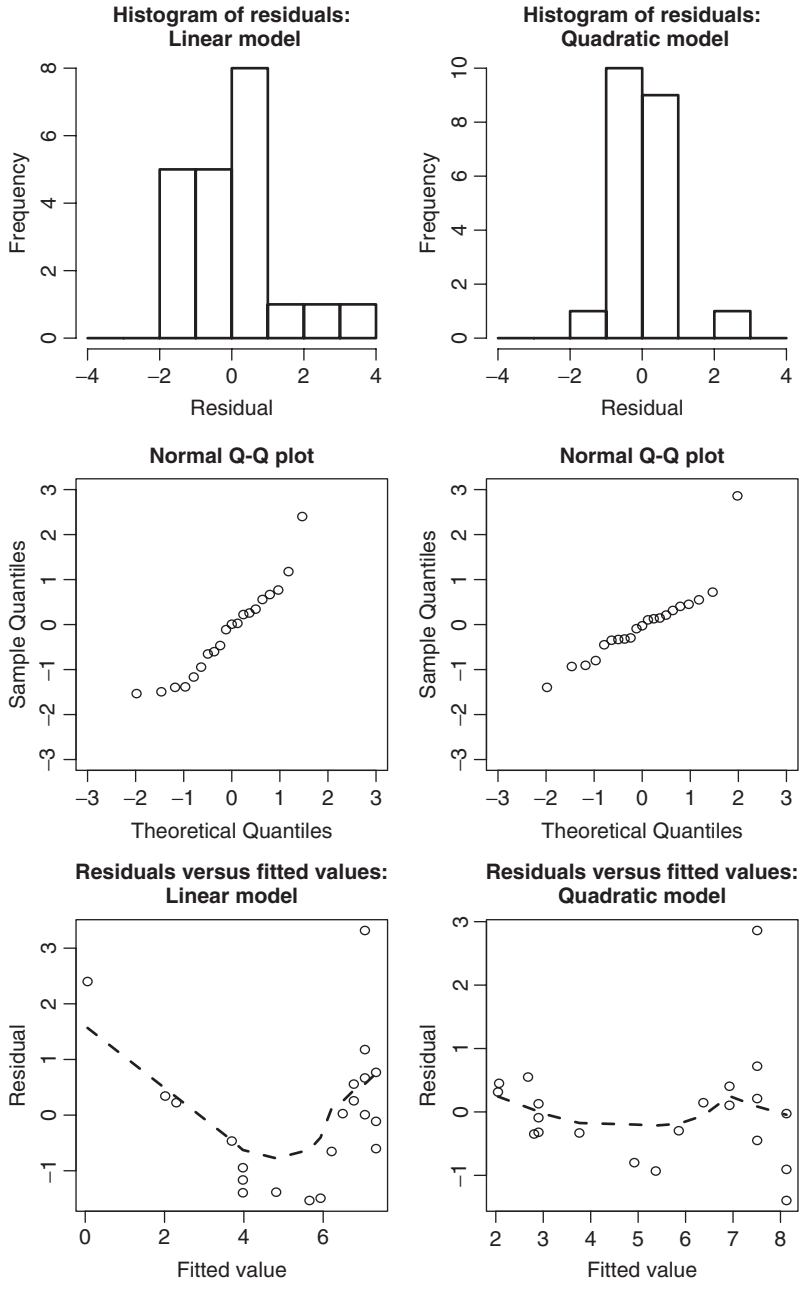
11.5.2 Residual diagnostics

Methods for checking the assumptions of linear regression models are based on the residuals, defined as the differences between the observed responses and the corresponding fitted values. The uses of the residuals for assessing normality, constant variance, and mean model misspecification are illustrated below using the fluoride data. Note that the independence assumption is not checked by examination of residuals, but rather should be assessed through consideration of the study design and data structure. For example, multiple observations per patient should always be assumed to be dependent.

Normality. The distribution of the residuals is used to check the normality assumption on the errors. For example, the histogram of the residuals from the linear model for the fluoride data (Figure 11.6, first row) exhibits skewness, which reflects fitted values that are a little too large for most of the observations and much too small for a few cases. In contrast, residuals from the quadratic model are close to symmetrically distributed. The normality assumption is also checked using the **Q-Q plot** (Figure 11.6, second row), which plots the residuals against the expected values of the residuals assuming they came from a normal distribution (Chapter 10). Residuals having an approximate normal distribution will exhibit an approximately straight line in the Q-Q plot. The fluoride residuals from the linear model exhibit clear lack of normality while those from the quadratic model fall roughly along a straight line with the exception of one outlier. With a sample size as small as 21 one can only detect very large departures from normality; thus, we would conclude that there is no evidence for nonnormality of the errors for the quadratic model, although the observation with the outlying residual would need to be investigated to determine if it is a valid value.

Constant variance. One way in which the constant variance assumption may fail is that there is a relationship between the mean and the variance of the responses.

[1] Stepwise regression is a class of methods for selecting covariates to be included in a regression model, usually based on statistical significance of the model terms. Although such methods are dangerous when applied to studies of causal relationships, they can be useful for creating models for prediction.


Histogram of residuals:
Linear model
Frequency
Residual
Histogram of residuals:
Quadratic model
Frequency
Residual
Normal Q-Q plot
Sample Quantiles
Theoretical Quantiles
Normal Q-Q plot
Sample Quantiles
Theoretical Quantiles
Residuals versus fitted values:
Linear model
Residual
Fitted value
Residuals versus fitted values:
Quadratic model
Residual
Fitted value

Such a mean-variance relationship may be detectable in a plot of the residuals versus the fitted values. The plots for the linear and quadratic models fitted to the fluoride data (Figure 11.6, third row) show a slight tendency for the spread of the residuals to increase with increasing fitted values. This pattern is commonly exhibited by positive-valued variables, such as DMFT. Two approaches are possible for dealing with a **mean-variance relationship**: (1) use robust 'sandwich' standard errors (see Chapter 13); or (2) apply a transformation of the response which eliminates the mean-variance relationship. A transformation should only be considered if the resulting model answers a scientific question of interest. For example, a logarithmic transformation of DMFT may eliminate the mean-variance relationship, but would not address an appropriate scientific question if a researcher wants to learn about mean DMFT. For example, estimating the total number of filled teeth in a population requires extrapolation from a mean for the sample; results based on log-transformed DMFT would not be useful. When transformations are not appropriate, robust standard errors (Chapter 13) can be used to obtain valid inferences about means even in the presence of nonconstant variance.

A logarithmic transformation is often applied to concentration data to eliminate a positive mean-variance relationship. For example, DeRouen *et al*. (2006) analyzed log-transformed urinary mercury concentrations, using regression models with the log-concentration of creatinine as one covariate. In such a model, a given percentage increase in creatinine is associated with a certain percentage increase in mercury concentration. A linear model for nontransformed concentrations is less meaningful in this setting.

Correctly specified mean model. In observational studies, it is nearly impossible to rule out a possible misspecification of the model for the mean response, because there could be confounders that are not known to the researcher. However, it is possible to explore the possibility of omitted variables using the data at hand. For example, a plot of the residuals versus fitted values from the linear model for the fluoride data (Figure 11.6, third row) strongly suggests a misspecification of the model because of the strong trend in the residuals displayed by the smooth curve. The corresponding plot obtained from the quadratic model shows little to no trend in the residuals, which suggests that the quadratic model fits reasonably well. Of course, this does not rule out the possibility of other variables being important predictors of DMFT.

Figure 11.6 Residual plots for the fluoride data based on a linear model (left-hand panels) and a quadratic model (right-hand panels). Histograms of residuals (first row) exhibit no clear outliers, but the distribution is somewhat skewed for the linear model. Q-Q plots of residuals (second row) show evidence of nonnormality for the linear model, but reasonable fit for the quadratic model with the exception of one outlier. Plots of residuals versus fitted values with smooth curves (third row) show clear misspecification of the linear model, but no clear departures from the quadratic model.

Table 11.3 Results from fitting straight-line models to the fluoride data with the community having the highest fluoride concentration (community #2, concentration 2.6 ppm) included or excluded.

	Community #2 Included		Community #2 Excluded	
Variable	estimate	SE	estimate	SE
(Intercept)	7.332	0.390	7.607	0.352
F	−2.797	0.386	−3.451	0.411

11.5.3 Influential observations

It is common in applications of regression for some observations to have more influence over the results than others. This is not in itself a bad thing but investigation of the validity and accuracy of such values is advisable. For example, in the fluoride data set the community with the highest fluoride level represents an influential observation. Excluding this observation from the analysis results in large differences in results when fitting the simple linear regression model (Table 11.3). This example illustrates the fact that observations with extreme values of one or more explanatory variables are likely candidates for influential observations because the model that best fits the other data points may not pass near this point, which implies that the model will need to change in order to reduce the residual for this point. Useful tools for screening data sets for influential data points are **Cook's distance** (Cook & Weisberg, 1982) and **DFBETA** (Belsley, Kuh & Welch, 1980).

11.6 Advanced regression methods

Linear regression is a very powerful technique for describing associations between variables. However, there are situations for which linear regression is not adequate. Firstly, relationships between quantitative variables may be inherently nonlinear, and it may or may not be possible to find transformations of the response or explanatory variables which induce a linear relationship. A second type of situation in which a non-linear model may be needed arises often in epidemiology, when the response variable is a discrete variable, such as a binary indicator variable of whether or not a patient has a particular disease of interest. The class of Generalized Linear Models (GLMs) was developed for this type of situation. Finally, when the response variable is the time until some event of interest, such as the survival time of a patient, then a different type of nonlinear regression model is needed.

11.6.1 Nonlinear regression

In pharmacokinetic studies, the concentration of drug in a patient (y) may be described as a function of time (t) using a model such as

$$y = \beta_1 \exp(-\beta_2 t) + \beta_3 \exp(-\beta_4 t) + \epsilon.$$

This is a nonlinear model because the parameters β_2 and β_4 are involved in the (nonlinear) exponential function. Specialized methods are required to fit such models (Seber & Wild, 1989; Bates & Watts, 2007).

11.6.2 Generalized linear models

A **generalized linear model (GLM)** (McCullagh & Nelder, 1989) takes the form

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{11.4}$$

Note that the right-hand side of the model equation is identical to that of a linear regression model, but the left hand-side includes a transformation of the mean response μ expressed as $g(\mu)$. The transformation g is called the **link function**.

One example of a GLM is **logistic regression**, which was developed for analysis of a binary response variable. A typical application of logistic regression is to the analysis of case-control studies, which aim to assess associations between putative risk factors and the binary indicator variable representing the presence or absence of a certain disease. In this setting, a logistic model describes the probability, p, that a patient is a case (rather than being a control), as a function of risk factors $x_1, \ldots, x_p$ using the following mathematical equation:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{11.5}$$

Comparing this to the general form of the GLM given above, we see that the link function for the logistic regression model is the **'logistic' transform** of the probability p, given by $\log(p/(1-p))$. Note that the mean is represented by p rather than μ simply because for a binary indicator variable the mean is a probability. For interpretation of logistic regression results, the estimates of the coefficients are exponentiated and interpreted as **odds-ratios**. For example, $\exp(\beta_1)$ represents the ratio of odds of being a case for two subjects with values of the factor x_1 differing by 1, with all other variables, $x_2, \ldots, x_p$ being equal.

A related type of regression analysis is **ordinal regression**, which is useful for response variables measured on an ordinal scale with categories arranged in a natural order. For example, caries scores are often ordinal scales (Ismail *et al*., 2007). A regression model in this situation must account for the categorical nature of the data, the existence of more than two possible levels, and the ordering among the levels. One useful model for this situation is the *proportional odds* model, which is based on a series of logistic regression models defined in terms of dichotomization of the ordinal variable using various cut-points.

A third type of GLM is the *log-linear* model for count data (also called *Poisson regression*). This model uses the logarithm as the link function and has the form $\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. As for logistic regression, the coefficients in a log-linear model are exponentiated for ease of interpretation. In this case, the exponentiated coefficient $\exp(\beta_1)$ represents the ratio of the mean number of events for two patients with values of the factor x_1 differing by 1, with all other variables, $x_2, \ldots, x_p$ being equal.

11.6.3 Survival analysis

In some clinical studies, interest lies in the time elapsed until an event of interest, such as failure of a dental implant. *Cox regression* is a method commonly applied in this situation. A Cox model is described in terms of the *hazard function* $\lambda(t)$, which is the instantaneous probability of the event occurring at time t, given that the event has not occurred prior to time t. The model takes the form

$$\log(\lambda(t)) = \log(\lambda_0(t)) + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{11.6}$$

where the term $\lambda_0(t)$ is the 'baseline' hazard function and assumes the role of the intercept, namely the hazard function for patients with the value 0 for all explanatory variables. In Cox regression, a regression coefficient is exponentiated and interpreted as a ratio of hazard rates. Typically, some events will be '**censored**' (Chapter 15), which means that they are not observed to occur prior to the end of the study or end of follow-up of a particular patient. Cox regression analysis accommodates censoring naturally, although it is necessary to make strong assumptions about the censoring process in order to obtain valid inferences.

11.7 Summary

This chapter presents the basic statistical concepts underlying the study of associations between variables. These concepts range from descriptive tools such as the scatterplot and the correlation coefficient to the inferential methods of regression analysis. Regression consists of a large class of methods for describing the association between a set of explanatory variables and a response variable. Different types of regression have been developed for handling response variables of different forms, including quantitative, binary, ordinal, count, and failure-time variables. A unifying feature of all regression methods is the common form for the right-hand side of the model, which is a linear combination of the predictor variables. Because of this common feature, many features of linear regression transfer with little modification to other types of regresion. For example, methods for coding explanatory variables to describe the treatment and covariate effects of interest apply to all types of regression model.

Departures from model assumptions can greatly impact the validity of the results of a regression analysis. In particular, misspecification of the mean model, for example because of omitted covariates, can lead to severely misleading inferences. All types of regression (not just linear regression) are susceptible to this problem. Correlation between responses is another type of model misspecification which can lead to seriously misleading inferences for all types of regression. The other types of misspecification considered here, non-constant variances and non-normal distributions, are less serious, and have different implications for different types of regression. Assumptions related to missing data, not considered in this chapter, are very important, and will be discussed in Chapter 14.

References

Bates, D. M. & Watts, D. G. (2007) *Nonlinear Regression Analysis and its Applications*. John Wiley & Sons, Inc., New York.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, Inc., New York.

Chilton, N. W. (1967) *Design and Analysis in Dental and Oral Research*, Philadelphia, Lippincott.

Cook, R. D. & Weisberg, S. (1982) *Residuals and Influence in Regression*. New York, Chapman & Hall.

DeRouen, T. A., Martin, M. D., Leroux, B. G., *et al*. (2006). Neurobehavioral effects of dental amalgam in children: a randomized clinical trial. *Journal of the American Medical Association* **295**: 1784–92.

Draper, N. R. & Smith, H. (1998) *Applied Regression Analysis* (3rd ed.). John Wiley & Sons, Inc., New York.

Ismail, A. I., Sohn, W., Tellez, M., *et al*. (2007). The international caries detection and assessment system (ICDAS): an integrated system for measuring dental caries. *Community Dentistry and Oral Epidemiology*, **35**: 170–8.

Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002) The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, **23**: 151–69.

McCullagh, P. & J. A. Nelder. (1989) *Generalized Linear Models*, 2nd ed, Chapman & Hall, New York.

Seber, G. A. F. (1984) *Linear Regression*. John Wiley & Sons, Inc., New York.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. John Wiley & Sons, Inc., New York.

12

Assessing accuracy of oral health diagnostic tests

Todd A. Alonzo and Peter J. Giannini

12.1 Introduction

Many new diagnostic tests, screening tests, and biomarkers are being developed for the early detection and diagnosis of oral health conditions. Diagnostic tests can be measured on a binary, ordinal, or continuous scale. Examples of tests with binary or dichotomous results (i.e. positive or negative) include culture for the diagnosis of fungal infections resulting from *C. albicans* and oral brush biopsy for the detection of premalignant or malignant oral epithelial lesions. Interpretations of images for the presence of caries are usually based on the following 5-point ordinal confidence scale: 1 = caries definitely absent, 2 = caries probably absent, 3 = unsure if caries absent or present, 4 = caries probably present, and 5 = caries definitely present. Examples of tests measured on a continuous scale include dental caries diagnosis with a laser and salivary biomarkers for the diagnosis of colonies of microorganisms associated with dental caries (*Streptococcus mutans* and *Lactobacillus acidophilus*) as well as periodontal disease (*Porphyromonas gingivalis*).

Before diagnostic tests are implemented in practice, it is imperative that the diagnostic accuracy of the tests is assessed. The **accuracy** of a test is the test's ability to correctly discriminate among alternative states of health. For example, presence or absence of caries, cancer, or more generally disease. True disease status is determined by a **gold standard** (see Chapter 9) that is assumed to measure disease status without error. When it comes to diagnosing oral lesions, sometimes the gold standard, or most ideal method of determining the final diagnosis, is via a scalpel

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

biopsy. This is where a piece of tissue from the lesion is submitted in 10 % formalin and processed for histopathologic examination and the pathologist renders a diagnosis. Careful attention should be given to the selection of the gold standard because even small errors in the gold standard can result in imprecise estimates of the accuracy of diagnostic tests (1). Section 12.6 discusses methods for assessing accuracy when the gold standard is not perfect. Wenzel and Hintze (2) provide a nice discussion on the choice of gold standard for evaluating tests for caries diagnosis.

In this chapter we will first describe different measures of test accuracy and methods for estimating accuracy. Then issues regarding study design will be discussed followed by methods for comparing test accuracy. Then issues with correlated diagnostic test data are discussed. The chapter will end with a discussion of methods for assessing accuracy when there is an imperfect gold standard and there is incomplete disease verification. For ease of presentation, diagnostic tests are usually referred to. However, the approaches and issues discussed are equally relevant for screening tests and biomarkers. Formulas are provided in this chapter for the interested reader but can be ignored without compromising the general understanding of the material.

12.2 Estimating accuracy for binary tests

A study was conducted to assess the diagnostic accuracy of the OralCDx technique, a computerized analysis of brush biopsies, to detect dysplasia or carcinoma in patients with oral mucosal lesions (3). The investigators were interested in quantifying the accuracy of brush biopsy results where brush biopsy results were classified as positive for dysplasia or oral squamous cell carcinoma or not. There are several measures of test accuracy that are available for diagnostic tests with binary results such as OralCDx. In this section we describe several measures of accuracy as well as approaches for estimating each measure.

12.2.1 Sensitivity and specificity

The most common measures of accuracy are sensitivity and specificity. **Sensitivity** is the proportion of diseased subjects that test positive by the diagnostic test. Table 12.1 summarizes 96 OralCDx results compared with the gold standard of histologic diagnosis. Of the 26 specimens determined to have disease, here dysplasia or carcinoma, by the gold standard, 16 tested positive with the OralCDx.

Table 12.1 Data from a study investigating the performance of OralCDx in detecting presence (D = 1) or absence (D = 0) of dysplasia or carcinoma in patients with oral mucosal lesions. Positive and negative OralCDx results are denoted $Y = 1$ and $Y = 0$, respectively.

	D = 1	D = 0	
$Y = 1$	16	2	18
$Y = 0$	10	68	78
	26	70	96

Therefore, the sensitivity, also referred to as true positive fraction (TPF), for OralCDx is estimated as the number of diseased specimens with a positive brush biopsy divided by the total number of diseased specimens or $16/26 = 0.615$. **Specificity** is the proportion of non-diseased subjects that test negative by the diagnostic test. Of the 70 specimens determined to not have dysplasia or carcinoma, 68 correctly tested negative with OralCDx. Thus, the specificity for OralCDx is estimated as 68/70 or 0.971. Specificity is equivalent to 1 minus the false positive fraction (FPF) which would be estimated as $1\text{-}0.971 = 0.029$ for OralCDx. The estimators of TPF and FPF are proportions so confidence intervals can be constructed using standard approaches appropriate for binomial proportions (see Chapter 10).

A perfect test has sensitivity and specificity both equal to 1 while a non-informative test is such that sensitivity is equal to 1-specificity or equivalently that TPF equals FPF. TPF quantifies the key benefit of screening, i.e. disease detection, while the FPF quantifies a key disadvantage of screening because subjects that have false positive results are sent for work-up procedures or treatments that are often costly in both human and monetary aspects. Therefore, when comparing tests, it is important to consider how tests compare in regards to both TPF and FPF.

12.2.2 Predictive values

The predictive value of OralCDx may also be of interest. Conversely to true and false positive fractions which quantify how well the test reflects true disease status, the predictive value of a test is the probability a subject has the disease given the results of a test. Specifically, **positive predictive value** (PPV) is the probability of disease in those with a positive test result. Consider again the OralCDx data in Table 12.1. To estimate the PPV of OralCDx, the number of specimens that are OralCDx positive and have dysplasia or carcinoma is divided by the total number of specimens that are OralCDx positive. Of the 18 specimens that tested positive with OralCDx, 16 had dysplasia or carcinoma. Thus, PPV is estimated as $16/18 = 0.89$. **Negative predictive value** (NPV) is the probability of not having the disease when the test result is negative. The NPV of OralCDx is estimated as the number of specimens without dysplasia or carcinoma that tested negative with OralCDx divided by the total number of specimens that tested negative with OralCDx. That is, NPV is estimated as $68/78 = 0.87$. Confidence intervals for PPV and NPV can be obtained using standard approaches for binomial proportions.

A perfect test has PPV and NPV both equal to 1 while a non-informative test has PPV equal to the population prevalence of disease, ρ, and NPV equal to $1 - \rho$. PPV and NPV depend not only on the TPF and FPF of the test but also on the prevalence of disease in the population in which the test is performed. Specifically,

$$PPV = \frac{TPF \times \rho}{TPF \times \rho + FPF \times (1 - \rho)} \quad \text{and}$$

$$NPV = \frac{(1 - FPF) \times (1 - \rho)}{(1 - FPF) \times (1 - \rho) + (1 - TPF) \times \rho}.$$

These previous expressions are derived in Chapter 18. Clearly, settings with low disease prevalence can yield low PPV even for tests with good TPF and FPF; whereas, settings with high disease prevalence can yield low NPV for tests with good TPF and FPF.

12.3 Estimating accuracy for nonbinary tests

Consider a study to determine the ability of two salivary biochemical markers measured on a continuous scale to accurately detect the presence of periodontitis. The accuracy of these markers could be assessed using the approaches discussed in the previous section if one was able to dichotomize the continuous results as above or below a specified threshold or cutpoint. Often it is not possible or of interest to dichotomize continuous results, so we next discuss a measure for assessing the accuracy of a marker or diagnostic test that yields continuous results.

12.3.1 ROC curve

Receiver operating characteristic curves (ROC curves) are a common measure of the accuracy of tests with continuous or ordinal results. For illustrative purposes, continuous tests are considered here. An **ROC curve** is a plot of the TPF versus FPF associated with all binary tests that can be formed by varying the cutpoint used to define a positive result.

ROC curves measure the amount of separation between the distribution of test results in the diseased population from the distribution of test results in the non-diseased population (Figure 12.1). When the distributions of test results for the diseased and non-diseased populations completely overlap, then the ROC curve is the 45 degree line from (0,0) to (1,1) indicating a non-informative test. The more separated the distributions, the closer the ROC curve is to the upper left-hand corner. A curve that reaches the upper left corner corresponds to a perfect test. ROC curves have the nice feature that they can be used to visually compare the accuracy of different tests even when tests are measured either in different units or on completely different scales.

12.3.2 Area under the ROC curve

The accuracy of a continuous test can be summarized by calculating the area under the ROC curve (AUC). AUC varies from 0.5 for an uninformative test to 1 for a perfect test. AUC is 0.56, 0.76, and 0.92 for the ROC curves in Figures 12.1d, 12.1e, and 12.1f, respectively. The AUC can be interpreted as the probability that the test result for a randomly chosen diseased subject exceeds that for a randomly chosen non-diseased subject (4). AUC can also be interpreted as the average TPF across all values of FPF.

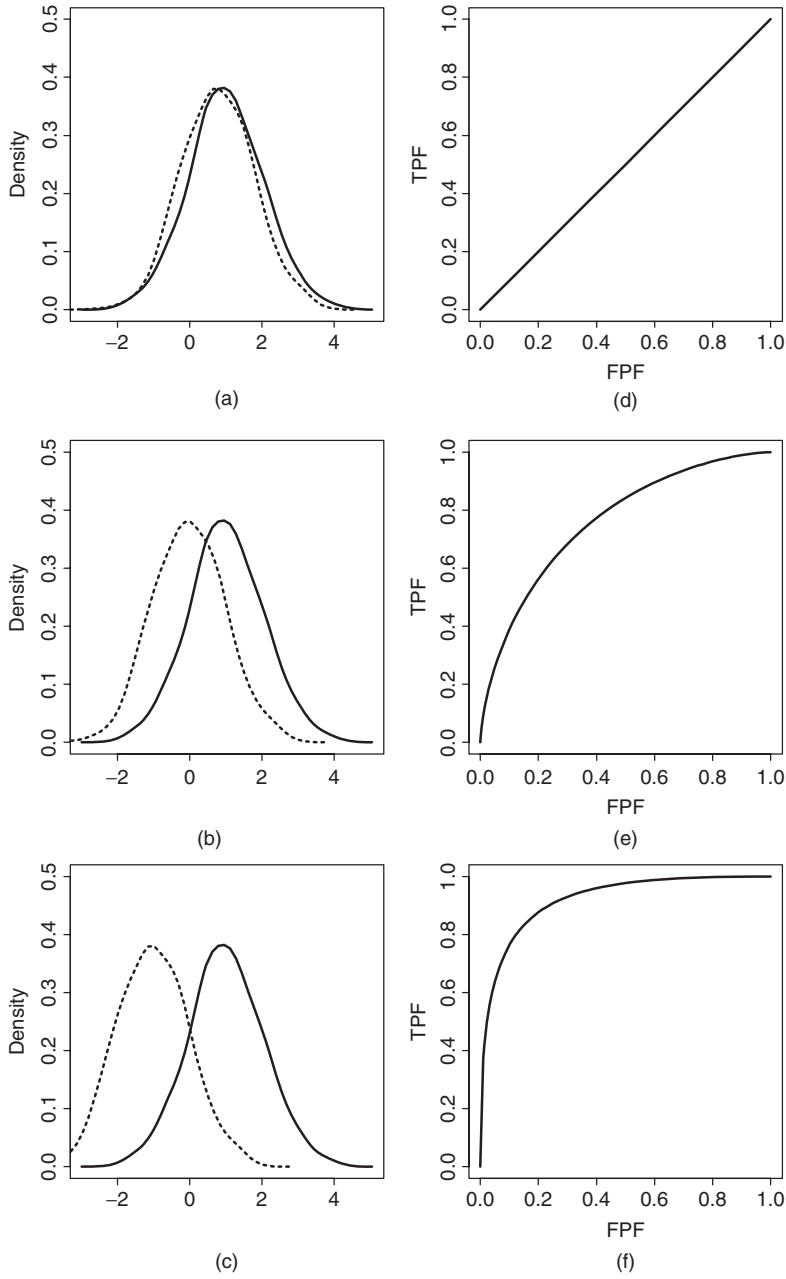


Figure 12.1 (a)-(c) Distributions of test results for diseased population (right curves) and non-diseased populations (left curves). Diseased populations correspond to a normal distribution with mean 1 and standard deviation 1. Non-diseased populations correspond to a normal distribution with standard deviation 1 and mean 0.8 for (a), 0 for (b), and −1 for (c). Corresponding ROC curves are given in (d)-(f).

12.3.3 Estimation

Approaches for estimating ROC curves and corresponding AUC summary statistics differ in the assumptions they make. Empirical and binormal approaches will be discussed. Empirical approaches allow the data to speak for themselves; whereas, binormal approaches make parametric assumptions. Let Y_{Di}, $i = 1, \cdots, n_D$ and Y_{Dj}, $j = 1, \cdots, n_{\bar{D}}$ be the continuous test results for the diseased and non-diseased subjects, respectively.

12.3.3.1 Empirical ROC curve

The empirical approach makes no assumptions about the distribution of the data, i.e. non-parametric. Specifically, this approach estimates the ROC curve using the observed estimates of $TPF(c)$ and $FPF(c)$ for all cutpoints c. Let $\widehat{TPF}(c)$ and $\widehat{FPF}(c)$ be the proportion of diseased and non-diseased subjects, respectively, with test results at or exceeding c. The empirical ROC curve is created by plotting $\widehat{TPF}(c)$ versus $\widehat{FPF}(c)$ for each cutpoint. Two salivary biochemical markers (Markers 1 and 2) were measured for 75 subjects with periodontitis and 75 subjects without periodontitis. The empirical ROC curves for the two markers (Figure 12.2)

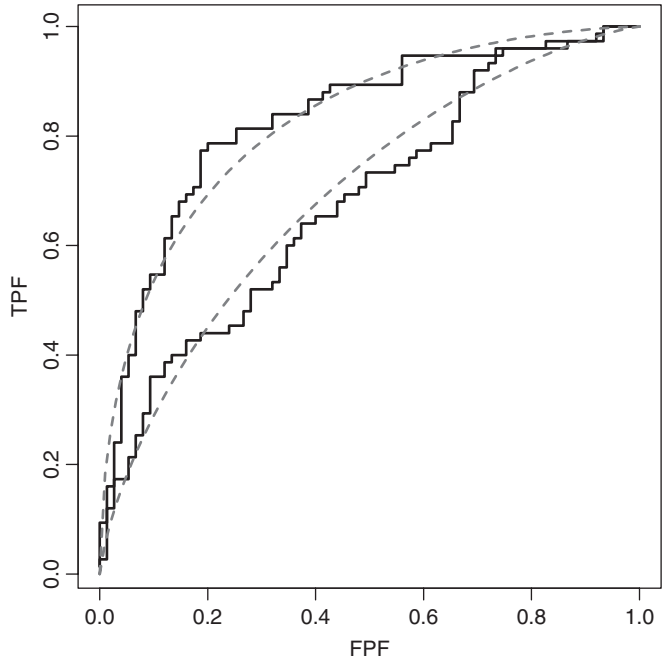


Figure 12.2 ROC curve for salivary marker 1 (top 2 curves) and marker 2 (bottom 2 curves). Jagged curves were estimated using the empirical approach while smooth curves were estimated using the binormal approach.

indicate that marker 1 does a better job of discriminating between those with and without periodontitis.

Extrapolation of the empirical ROC curve to all possible cutpoints can be made by connecting observed data points linearly. For data with no ties (when no two data values are the same), adjacent points are connected with horizontal and vertical lines resulting in a step function. As the threshold changes, inclusion of a true positive result produces a vertical jump of size $1/n_D$ and inclusion of a false positive result produces a horizontal jump of size $1/n_{\bar{D}}$. When there are ties in the data between diseased and non-diseased test results, both the true positive and false positive fractions change simultaneously, resulting in a point displaced both horizontally and vertically from the last point.

The AUC can be estimated as the area under the empirical ROC curve. It can be shown that this is equal to

$$\widehat{AUC}_e = \sum_{j=1}^{n_{\bar{D}}} \sum_{i=1}^{n_D} \left\{ I[Y_{Di} > Y_{\bar{D}j}] + \frac{1}{2} I[Y_{Di} = Y_{\bar{D}j}] \right\} / n_D n_{\bar{D}}, \tag{12.1}$$

where $I[\ ]$ is equal to 1 (0) if the expression is true (false). Using this equation, estimates of AUC are 0.83 and 0.68 for salivary biochemical markers 1 and 2, respectively.

Equation 12.1 is equal to the Mann-Whitney two-sample statistic for comparing the distributions of test results in the diseased and non-diseased populations (4). When there are no tied data points, the empirical AUC is calculated by comparing each disease test result with each non-disease result and determining the percentage of the time that the disease test result is larger than the non-disease test result. Variance expressions for $\widehat{AUC}_e$ are available for independent (5) and correlated or clustered (6) test results. See Section 12.5 for a discussion of situations with correlated results.

12.3.3.2 Binormal ROC curve

The binormal approach assumes that the ROC curve has a particular form, namely a binormal form. More specifically, if the test results from the diseased population and the test results from the non-diseased population have normal distributions with means μ_D and $\mu_{\bar{D}}$ and standard deviations σ_D and $\sigma_{\bar{D}}$, then the corresponding ROC curve has the binormal function form:

$$ROC(t) = \Phi(a + b\Phi^{-1}(t)), \tag{12.2}$$

where t is the FPF on the x-axis, $ROC(t)$ is the TPF on the y-axis, Φ is the standard normal cumulative distribution function, a is the intercept, and b is the slope. This binormal model only requires that there exists a monotone transformation (for example, log, negative, or square root transformation) of the data that will result in normally distributed diseased and non-diseased test results.

Using the binormal model, the AUC can be estimated using

$$\widehat{AUC} = \Phi\left(\frac{\hat{a}}{(1+\hat{b}^2)^{\frac{1}{2}}}\right). \tag{12.3}$$

Wieand *et al*. (7) provides an expression for the variance of this estimate.

Different approaches exist for estimating a and b. One approach assumes that the test results from the diseased population are normally distributed with mean μ_D and standard deviation σ_D and the test results from the non-diseased population are normally distributed with mean $\mu_{\bar{D}}$ and standard deviation $\sigma_{\bar{D}}$. Then a can be estimated as $(\hat{\mu}_D - \hat{\mu}_{\bar{D}})/\hat{\sigma}_D$ and b can be estimated as $\hat{\sigma}_{\bar{D}}/\hat{\sigma}_D$. Other approaches assume that there exists some unknown transformation that makes the test results normally distributed (8, 9).

Once again consider the periodontitis data (Figure 12.2). For the subjects with periodontitis, the mean and standard deviation are 1.347 and 0.979 for marker 1 and 0.575 and 0.976 for marker 2. For the subjects without periodontitis, the mean and standard deviation are 0.072 and 0.928 for marker 1 and −0.112 and 0.961 for marker 2. Assuming that the marker values are normally distributed, a is estimated to be $(1.347 - 0.702)/0.979 = 1.302$ for marker 1 and 0.704 for marker 2. Furthermore, b is estimated to be $0.928/0.979 = 0.948$ for marker 1 and 0.985 for marker 2. The smooth curves in Figure 12.2 are obtained by inserting the estimates of a and b into Equation 12.2. Using Equation 12.3, estimates of AUC for these smooth ROC curves are 0.828 and 0.692 for markers 1 and 2, respectively. These estimates are similar to the estimates previously obtained using the empirical approach.

12.4 Comparison of accuracy

This section presents a discussion of study design issues and approaches for comparing TPF and FPF for two binary tests (Section 12.4.2) and two ROC curves for a continuous test (Section 12.4.3).

12.4.1 Study design issues

Studies to compare the accuracy of diagnostic tests can be performed prospectively or retrospectively. Retrospective studies involve selecting subjects on the basis of their true disease status, as determined by the gold standard, and performing the tests on them. These retrospective studies are often called case-control studies where cases are those with disease and controls are those without disease. Prospective studies involve applying the tests to a random sample from the population of interest and determining true disease status for all study subjects.

Studies designed to compare the accuracy of multiple diagnostic tests can have a paired or unpaired design. In a paired design all tests are performed on each individual. Conversely in an unpaired design each individual is only administered one of the tests. Pairing is often desirable because it can reduce variability in making comparisons between tests by eliminating between-subject variance. Therefore,

pairing is usually a more efficient design requiring smaller sample sizes. However, if the adminstration of one test interferes with the results of another test, an unpaired design may be necessary.

12.4.2 Binary tests

To compare the accuracy of two binary tests, the absolute difference, odds ratio, or ratio of TPFs or FPFs for the two tests can be used. The ratio is presented here because it has a straightforward interpretation and also has advantages in that statistical inference on the relative scale is less difficult than inference on the absolute scale, and the interpretation on the relative scale is less awkward than that for odds ratios (1). Consider the relative true positive fraction rTPF(A,B) = TPF_A/TPF_B and relative false positive fraction rFPF(A,B) = FPF_A/FPF_B, where subscripts denote the diagnostic test. Specifically, the null hypothesis H_0: $TPF_A = TPF_B$ is equivalent to H_0:rTPF(A,B) = 1 so it can be concluded that the tests have different TPF or FPF if the confidence intervals for the ratios do not contain the value 1. For both designs rTPF and rFPF can be estimated as the ratio of TPF and FPF estimates for the two diagnostic tests. Different approaches are required to calculate confidence intervals for studies with paired and unpaired designs. Next, comparison of two binary diagnostic tests are illustrated using a paired design and then an unpaired design.

12.4.2.1 Paired design

A study was performed to determine the accuracy of salivary biochemical markers to detect periodontal disease, as determined by periodontal pocket probing (10). This is a paired design since lactate dehydrogenase (LDH) and alkaline phosphatase (ALP) were both assessed on all 187 study subjects as well as definitive disease assessment. Table 12.2 contains LDH and ALP results and whether gingivitis or periodontitis was present. The number of concordant and discordant LDH and ALP results have been simulated since they were not provided in Nomura *et al.* (10). Based on the data in Table 12.2, the estimate of TPF, $\widehat{TPF}$, for LDH is 82/124 = 0.661 and $\widehat{TPF}$ for ALP is 67/124 = 0.540 so that $\widehat{rTPF}(LDH, ALP)$ is 1.22 with 95 % confidence interval (1.00, 1.50) constructed using the approach of Cheng and Macaluso (11). These results suggest that LDH detected 22 % more

Table 12.2 Results of LDH and ALP for detecting the presence (D = 1) or absence (D = 0) of gingivitis or periodontitis. LDH (ALP) equal 1 corresponds to a positive LDH (ALP) test result; whereas, LDH (ALP) equal 0 corresponds to a negative LDH (ALP) test result.

	D = 1			D = 0		
	ALP = 1	ALP = 0		ALP = 1	ALP = 0	
LDH = 1	45	37	82	18	3	21
LDH = 0	22	20	42	12	30	42
	67	57	124	30	33	63

cases of gingivitis or periodontitis than ALP and with 95 % confidence the increase in detection is between 0 % and 50 %. $\widehat{FPF}$ for LDH is 21/63 = 0.333 and $\widehat{FPF}$ for ALP is 30/63 = 0.476 so that $\widehat{rFPF}(LDH,ALP)$ is 0.70 with 95 % confidence interval (0.52, 0.95). Therefore, LDH detected 30 % ((1-0.7) times 100 %) fewer subjects without periodontal disease as having disease than ALP and with 95 % confidence the decrease in detection is between 5 % and 48 %.

12.4.2.2 Unpaired design

We previously compared the accuracy of two salivary markers, LDH and ALP, for detecting periodontal disease. Free hemoglobin (f-Hb) levels were also collected in this study. For illustrative purposes, Table 12.3 presents LDH and f-Hb results for those with and without gingivitis or periodontitis as if they were measured in different subjects (i.e. unpaired design). Using these data, we calculate that the $\widehat{TPF}$s for LDH and f-Hb are 82/124 = 0.66 and 33/122 = 0.27 so that $\widehat{rTPF}(LDH$, f-Hb) is 2.44 with a 95 % confidence interval (1.78, 3.36) calculated using the variance expression provided in Pepe (1). Similarly, we calculate that the $\widehat{FPF}$s for LDH and f-Hb are 21/63 = 0.33 and 12/63 = 0.19 so that $\widehat{rFPF}$(LDH, f-Hb) is 1.75 with 95 % confidence interval (0.94, 3.24). Therefore, we conclude that LDH detects 2.44 times more cases of periodontal disease than f-Hb but 1.75 times more subjects free of periodontal disease test positive with LDH than f-Hb.

12.4.3 Continuous tests

ROC curves for two diagnostic tests measured on continuous scales can be compared by comparing the corresponding AUC estimates. The null hypothesis of equal empirical ROC curves for tests A and B can be tested by comparing $\Delta\widehat{AUC}_e/\text{var}(\Delta\widehat{AUC}_e)^{\frac{1}{2}}$ with a standard normal distribution, where $\Delta\widehat{AUC}_e$ is the difference in the empirical estimates of AUC for tests A and B. If the two ROC curves are estimated using an unpaired design, then the variance of $\Delta\widehat{AUC}_e$ can be estimated as the sum of the variances for the AUC estimates for tests A and B. If the two ROC curves are estimated using the same set of study subjects (i.e. a paired design), then the correlation between the AUC estimates must be taken into account (1).

Two binormal ROC curves can be compared by comparing the estimated intercept and slope parameters for the two curves (12) or by comparing the estimated AUC for the two tests (7).

Table 12.3 Results of LDH and f-Hb for detecting the presence (D = 1) or absence (D = 0) of gingivitis or periodontitis. Y equal 1 corresponds to a positive LDH or f-Hb test result; whereas, Y equal 0 corresponds to a negative test result.

	D = 1			D = 0		
	Y = 1	Y = 0		Y = 1	Y = 0	
LDH	82	42	124	21	42	63
f–Hb	33	89	122	12	51	63

12.5 Issues with correlated diagnostic test results

Multiple test measurements are often made on the same individual in studies of the accuracy of oral health diagnostic tests. These studies result in correlated data that must be properly accounted for in the data analysis. In this section we will discuss a few different types of studies that result in correlated data.

We introduced in Section 12.4.1 the idea of studies with a paired design in which multiple diagnostic tests are performed on each individual. Results of the diagnostic tests are correlated in these studies because they are measured on the same individual. The analytic methods described in Section 12.4 take into account the correlation for the test results on each individual.

Correlated data can also arise in studies where the same diagnostic test is measured multiple times on an individual. This occurs, for example, in studies where multiple surfaces of a tooth (e.g. proximal, occlusal) are tested for carious lesions or multiple teeth from the same individual are tested. These studies result in clustered data where results for each surface are clustered within a tooth and results for each tooth are possibly clustered within an individual. Several analytic methods exist to quantify and compare the accuracy of diagnostic tests when the data are correlated or clustered. For example, different regression techniques exist for tests measured on binary, ordinal, and continuous scales. See Pepe (1) and Zhou *et al*. (13) for descriptions of these regression techniques. Regression methods have the nice feature that they can also be used to determine the effects of other factors on test accuracy. Examples of factors that could influence test accuracy include attributes of the individuals tested (e.g. age, gender, race), characteristics or severity of their disease (e.g. histopathologic grade and stage of cancer), and conditions under which the diagnostic tests are studied. It may be important to identify and understand the influence of these factors because populations and settings where a test is more or less accurate can be identified, which can be useful in determining how best to use a test.

Studies in which multiple observers, also referred to as readers, each read or interpret results (often resulting from some form of imaging modality) from the same individuals also yield correlated data. Correlation results from the fact that each observer has provided their subjective interpretation of test results for all individuals. Within and between observer agreement can be assessed using kappa values (see Chapter 9) when results are binary, ordinal, or nominal. Reliability can measured using the intra-class correlation coefficient (see Chapter 13) for continuous results. If disease ascertainment is also available, then the accuracy of the different observers relative to the disease status can be assessed. For example, Harase *et al*. (14) performed a study to compare the accuracy of proximal caries detection by extraoral tuned aperture computed tomography (TACT), intraoral TACT, and film radiographs. Seven observers scored 80 proximal surfaces for the presence or absence of proximal caries for the three imaging modalities using the 5-point confidence scale: 1 = caries definitely absent, 2 = caries probably absent, 3 = unsure if caries absent or present, 4 = caries probably present, and 5 = caries definitely present. Micro CT was used as the gold standard for determining presence of caries.

When observer readings are measured on a 5-point confidence scale, as is usually the case, the accuracy of diagnostic tests can be measured using the area under the ROC curve. To account for the correlated data, mixed-effects ANOVA models can be applied to the AUC for each combination of observers and tests (15).

12.6 Imperfect gold standard

The previous sections presented methods for assessing the accuracy of oral health diagnostic tests as compared to a perfect gold standard that measures disease without error. In practice, however, a perfect gold standard may not be available so an imperfect reference is used instead. An example of an imperfect reference may include performing cytology for the diagnosis of oral candidiasis instead of culture. With the erythematous form of candidiasis there may be too few hyphae to detect with cytology that otherwise might be detected by performing a culture on Sabouraud agar. A major drawback to assessing the accuracy of diagnostic test relative to an imperfect reference standard is that estimates of accuracy may be biased and the magnitude and direction of the bias is unknown. Adjustments for the bias can be made if the accuracy of the reference standard relative to a gold standard is known; however, that is usually not the case. Another approach is to combine several imperfect reference tests to form a composite reference standard that is potentially more accurate than each reference test (16). This composite reference standard has the advantages that several sources of information are used to generate the standard and the standard is independent of the test being investigated; the drawback to the composite reference standard is that the standard is imperfect. An alternative approach is latent class analysis which assumes that there is an unobserved latent binary variable that indicates the presence or absence of disease and relates results of several diagnostic tests with a statistical model. Drawbacks to this approach include that incorrect model assumptions can yield biased estimates of accuracy, the model assumptions are not fully testable with observed data, and a formal definition of disease is not required (17). Chapter 16 discusses the impact of measurement error and misclassification as well as possible methods to account for it. Clearly, there are challenges to assessing accuracy when an imperfect gold standard is not available and alternative solutions are required.

12.7 Incomplete disease ascertainment

The methods described in this chapter for assessing accuracy of diagnostic tests assumed that disease status, e.g. presence or absence of cancer, was available for all study subjects. However, in practice the gold standard test may be too costly, e.g. administering an MRI for the screening of breast cancer, or invasive, e.g. scalpel biopsy, to administer to all study subjects. Therefore, disease status assessment may be more likely in higher risk subjects than subjects at lower risk. This selective disease verification can result in biased estimates of accuracy (verification bias (18)) if the estimation methods do not properly account for the non-random

disease ascertainment. Methods for estimating accuracy that properly account for the non-random disease ascertainment are available [e.g. (18, 19)] provided that the data on all the factors that affected selection for disease verification were collected. Development of methods that require less restrictive assumptions regarding the missing disease values is an area of active research.

12.8 Summary

This chapter reviewed methods for evaluating the ability of diagnostic tests, screening tests, and biomarkers to correctly distinguish between disease and no disease. Appropriate methods are predicated upon the scale on which the test results are measured. Issues with correlated data frequently encountered in oral health studies were also discussed. Stata and S-plus were used to analyze the data in this chapter. A list of available commercial and non-commercial software to perform many of the methods discussed in this chapter can be found online at http://www.fhcrc.org/labs/pepe/dabs. This website also contains other helpful resources.

References

[1] M.S. Pepe (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York, NY.

[2] A. Wenzel & H. Hintze (1999) The choice of gold standard for evaluating tests for caries diagnosis *Dentomaxillofacial Radiology* **28**, 132–6.

[3] C. Scheifele, A. Schmidt-Westhausen, T. Dietrich & P.A. Reichart (2004) The sensitivity and specificity of the OralCDx technique: evaluation of 103 cases *Oral Oncology* **40**, 824–8.

[4] D. Bamber (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph *Journal of Mathematical Psychology* **12** 387–415.

[5] J.A. Hanley & B.J. McNeil (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve *Radiology* **143**, 29–36.

[6] N.A. Obuchowski (1997) Nonparametric analysis of clustered ROC curve data *Biometrics* **53**, 567–78.

[7] S. Wieand, M.H. Gail, B.R. James & K.L. James (1989) A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data *Biometrika* **76**, 585–92.

[8] C.E. Metz, B.A. Herman & J.H. Shen (1998) Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data *Statistics in Medicine* **17**, 1033–53.

[9] M.S. Pepe (2000) An interpretation for the ROC curve and inference using GLM procedures *Biometrics* **56**, 352–9.

[10] Y. Nomura, Y. Tamaki, T. Tanaka, *et al*. (2006) Screening of periodontitis using salivary enzyme tests *Journal of Oral Science* **48**, 177–83.

[11] H. Cheng M. & Macaluso (1997) Comparison of the accuracy of two tests with a confirmatory procedure limited to positive results *Epidemiology* **8**, 104–6.

[12] C.E. Metz & H.B. Kronman (1980) Statistical significance tests for binormal ROC curves *Journal of Mathematical Psychology* **22**, 218–43.

[13] X-H Zhou, N.A. Obuchowski & D.K. McClish (2002) *Statistical Methods in Diagnostic Medicine* John Wiley & Sons, Inc., New York, NY.

[14] Y. Harase, K. Araki & T. Okano (2006) Accuracy of extraoral tuned aperture computed tomography (TACT) for proximal caries detection *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology* **101**, 791–6.

[15] N.A. Obuchowski (1995) Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations *Academic Radiology* **2**, S22–S29.

[16] T.A. Alonzo, M.S. Pepe & C.S. Moskowitz (2002) Sample size calculations for comparative studies of medical tests for detecting presence of disease *Statistics in Medicine* **21**, 835–52.

[17] M.S. Pepe & H. Janes (2007) Insights into latent class analysis of diagnostic test performance *Biostatistics* **8**, 474–84.

[18] C.B. Begg & R.A. Greenes (1983) Assessment of diagnostic tests when disease verification is subject to selection bias *Biometrics* **39**, 207–15.

[19] T.A. Alonzo & M.S. Pepe (2005) Assessing accuracy of a continuous screening test in the presence of verification bias *Applied Statistics* **54**, 173–90.

Part IV

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

13

Analysis of correlated responses

Melissa D. Begg

13.1 Introduction to issues in analyzing clustered data

Most statistical techniques are designed for analyzing independent data; that is, for simple random samples, in which each member of the population under study has an equal probability of being selected for the study sample. Clustered data are different, because the probability of selection is not the same for all members of the population. Study subjects (or observations) enter the study in groups or 'clusters,' pre-defined by some criteria, and so do not represent independent observations. Settings in which cluster-correlated data are common include longitudinal studies (when we observe the same individuals on multiple occasions over time), family studies (when we record information on multiple siblings in a family), group-randomized trials (where there are multiple patients within a clinic, centre, or classroom), and in oral health research (which typically looks at multiple sites within a single patient).

There are many reasons why researchers choose a clustered study design. Sometimes it is chosen in order to increase precision by making comparisons within cluster (thereby reducing the additional variance imparted by more heterogeneous comparisons). On other occasions, clustered designs are employed in order to reduce the potential for bias due to confounding by making comparisons within cluster (controlling for a host of secondary covariates, both measured and unmeasured). In the setting of longitudinal research, clustering allows us to

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

examine within-subject patterns over time. Finally, on some occasions, a clustered design is the only feasible approach. One example would be that of an educational intervention in a school-based setting, where students can only be instructed in pre-defined groups (classrooms). Another example, of greatest interest to the oral health researcher, is that of periodontal studies (see Chapter 22 for an example). In periodontal research, observations are recorded at numerous sites within the mouth of a single subject. These represent, of course, natural clusters of observations.

How these observations are incorporated into an oral health analysis depends on the nature of the outcome and predictor variables, the purpose of the study, and the investigators' hypotheses about the underlying nature of the disease process under study. For example, an oral health researcher may choose to summarize data at the subject-level by taking the mean or median of the many observations/sites within an individual. Alternatively, a subject might be classified as a 'case' (someone with a diagnosis of periodontal disease) or not, depending on some summary of the site-specific measurements. Or the investigator may choose instead to analyze the information recorded at each site (in a so-called 'site-specific' analysis), relating site-specific outcome information to site-specific or subject-level exposure information, possibly adjusting for site-specific or subject-level covariates to control for confounding. In addition, these outcome measurements (which might represent extent of disease, severity, or progression) can take the form of dichotomous, ordinal, or fully continuous random variables. There is clearly a wide variety of combinations of predictor and outcome variables, and the methods selected for analyzing the data will have to be tailored to the structure at hand.

When the outcome under consideration is measured at the level of the site, specialized analytic approaches must be used in order to preserve the validity of the resulting inferences. In particular, clustered data tend to yield correlated responses, because two observations (sites) within the same cluster (subject) tend to be more similar than two observations from two different clusters. This correlation, referred to as *intra-cluster correlation*, or *ICC*, must be accounted for in an analysis because it can affect the estimation of the variances of parameter estimates. The presence of ICC means that observations are not independent; hence standard analytic techniques are not valid, in the sense that test statistics and confidence intervals generated while ignoring the ICC may be incorrect (i.e. may have a type I error rate other than that specified, or coverage probabilities for confidence intervals that deviate from the desired level). These problems are not merely esoteric statistical concerns; they can lead to flawed conclusions regarding the efficacy of an intervention, mechanisms of disease, or identification of risk factors for disease progression.

This problem was identified in the late 1980s by dentists and statisticians working in oral health, who issued a warning to their colleagues to beware of the consequences of ignoring ICC. As stated by Fleiss, Park, and Chilton (1987), 'Given the sizable within-mouth correlations in this study, we conclude that the assumption made by some investigators of virtual independence between sites is incorrect and that statistical procedures in which sites are the unit of analysis may be invalid.' This concern was echoed by Fleiss *et al.* (1988): 'taking sites

rather than patients as the units of statistical analysis in comparative clinical trials will tend to produce underestimated standard errors and thus produce overstated statistical significance and unduly narrow confidence intervals.' Further research demonstrated that the degree of bias in estimating standard errors would depend upon two items: the cluster size (number of observations per cluster) and the ICC (the level of correlation or similarity between items within the same cluster). Bias in the standard errors from an analysis that ignores clustering will increase as either the cluster size or the ICC level increases. The distortion can go in either direction (leading to standard errors that are under-estimated or over-estimated), and the direction of the distortion may depend upon the nature of the exposure variable (Laster, 1992). In general, if the exposure variable is subject-specific (fixed within subject), this tends to lead to underestimation of standard errors in a 'naïve' analysis (i.e. one that ignores the clustering among observations), as noted by Fleiss *et al*. (1988). However, if the exposure is recorded at the level of the site within subject (and, therefore, varies within subject), then this tends to lead to overestimation of standard errors in analyses that disregard the clustering (Laster, 1992). Simulation studies have shown that the degree of bias can be substantial. For example, Begg and Paykin (2001) show that actual rejection rates over 500 simulated data sets for a subject-specific exposure and a nominal 5 % significance level can be as large as 20 %, 40 % or even 60 % (assuming cluster sizes ranging from 4 to 32 and ICC ranging from 0 to 0.5).

Given the importance of adjusting for clustering in oral health research, investigators have developed different approaches for analyzing oral health data. For example, some investigators choose to collapse the multiple observations on a subject into a single summary statistic (or 'whole-mouth' statistic), like the mean, median, or maximum value. While there are occasions where this is perfectly appropriate (i.e. this approach is consistent with the investigators' hypotheses and understanding of the underlying biology of the disease), there are also occasions on which this approach may be inappropriate or inefficient. Whole-mouth averages, for example, may conceal associations that exist at the site-level, or adjustment for confounding across the mouth may not be sufficient to control bias at the site level. Another consideration, though perhaps not as important as the previous two, is power, or the ability to detect a significant association between an exposure and response at a given sample size. Reducing the data by summarizing will, obviously, reduce the sample size and the corresponding power for detecting an association.

Throughout this chapter, we will consider methods that are appropriate for the analysis of clustered data. Most of these methods will be illustrated on the data set described below using STATA software. STATA commands will be provided along with the corresponding STATA output for illustration.

13.1.1 Example: oral health and HIV infection

Lamster *et al*. (1994) conducted an observational study of HIV-infected and uninfected subjects with varying degrees of periodontitis. In this study, information was recorded at the level of the subject and at the level of the site during periodontal

examination. One of their objectives was to assess the relationship between HIV status and local immune response (as measured by level of immunoglobulin M, or IgM) after adjustment for the severity of periodontal disease at that site (as measured by attachment level in millimeters). Note that this study was conducted prior to the era of highly-active anti-retroviral therapy, and so its conclusions may not generalize to HIV-infected populations today. The subsample presented here includes data from 29 men, 17 of whom are HIV-infected and 12 of whom are uninfected. The number of sites per subject ranges from 9 to 16, with a median of 13. The total number of sites over all 29 subjects is 382. In analyses later in this chapter, we will relate IgM level to HIV status, adjusting for the effects of attachment level.

13.2 Approaches to analyzing clustered data

As noted earlier, there are different approaches for analyzing clustered data, all of which try to address the problem of the correlation among observations. These include: summarization over the cluster, repeated measures Analysis of Variance (ANOVA), generalized linear models for discrete and continuous responses, and bootstrap resampling by cluster.

13.2.1 The summary statistic approach

The simplest strategy is to generate a summary statistic (like the mean) over all observations in the cluster. This reduces multiple responses to a single summary response for each subject. Then one may compare summary responses across subjects; because there is one summary for each subject, observations are independent and standard statistical methods can be applied. Sample summary measures include the mean, median, maximum value across sites, number of sites or teeth meeting some pre-specified criterion (e.g. number of sites with attachment level greater than or equal to 3 millimeters), or a binary summary measure (e.g. a diagnosis of periodontitis). The approach offers the advantages of being simple and intuitive, the ability to use standard statistical techniques for analysis, and conceptual appeal in certain settings. Disadvantages are, however, numerous. Averaging over many sites may result in decreased sensitivity to effects. Furthermore, this approach may not accurately reflect the research hypothesis relating a site-specific exposure to a site-specific outcome. There may also be a loss of power and precision, due to the decreased sample size. Finally, there is no way to control for confounding at the level of the site, which may be crucial in reducing bias.

13.2.2 Repeated measures ANOVA

A second strategy is to apply the repeated measures (RM) Analysis of Variance (ANOVA) method, a version of the two-way ANOVA approach introduced in Chapter 10, but where one factor is the subject (cluster) effect (assumed to be random), and the other is the treatment/exposure effect (considered fixed). This

approach assumes that the correlation among observations is exchangeable (that is, that the level of correlation between any two sites in the mouth is the same, regardless of the distance between them). Advantages of this approach include the ability to accommodate multiple sites per subject and the opportunity to obtain estimates of the correlation among multiple responses per subject. The disadvantages include the requirement that the outcomes are continuous in nature, the exposure is categorical, and the assumption that the correlation structure is exchangeable (which may or may not be the case). Because oral health response data can be discrete or continuous, as can exposure data, the RM ANOVA approach is arguably too limiting for a broad range of applications, so we consider more flexible approaches below.

13.2.3 Generalized linear models for clustered data

Generalized linear models (GLMs) for clustered data are more general than either of the two previously described approaches. GLMs can accommodate continuous or discrete outcomes for a variety of distributions, are appropriate for clusters of unequal sizes, and can allow for a more general correlation structure than many other methods. To introduce GLMs, we first need to define some notation, and we do so in the setting of independent data for simplicity (we will extend this to clustered data later). Suppose that the outcome for subject i is represented by Y_i, the exposure is denoted by x_i, and random error terms by ε_i. For example, in the ordinary linear model for continuous, normally distributed data, we assume that the response variable Y_i is distributed as $N(\mu_i, \sigma^2)$ and where $\mu_i = \beta_0 + \beta_1 X_i$. This is equivalent to assuming $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where the ε_i $(i = 1, \ldots, n)$ are distributed as independent, normal random variables with mean zero and variance σ^2. GLMs for independent data (where response variable Y_i is assumed to follow distribution $F(\mu_i, \sigma^2)$, where F could represent the normal distribution, the binomial, the Poisson, etc.) generalize this expression as follows:

$$h(\mu)_i = \beta_0 + \beta_1 X_i \tag{13.1}$$

where h(.) is called the 'link' function (see Chapter 11). Common choices for the link function and the data distribution include: identity link with normal errors as in linear regression; logit link with binomial errors as in logistic regression; and log link with Poisson errors as in Poisson regression.

We will discuss two forms of GLM for clustered data: a random effects model (in particular, we will focus on a random intercept model) and a marginal model (specifically, we will consider the Generalized Estimating Equation, or GEE, technique for clustered data). A key feature of the random effects model is that the *conditional* mean of the response variable Y, given a set of random effects, is modeled as a function of the covariates X and the random effects that may vary from one cluster to the next. The heterogeneity among clusters (or correlation within cluster) is assumed to be of interest, and so is explicitly modeled. For the marginal model, the marginal (as opposed to conditional) mean of the response variable Y is modeled as a function of covariates X with regression coefficients that are fixed

from cluster to cluster. The within-cluster correlation is modeled separately from the regression model for Y on X, and is treated as a nuisance parameter (not of primary interest). Below we refine notation and introduce the random effects and marginal approaches, applied to the HIV Study data.

13.2.3.1 The random effects approach

Let's set notation. Index $i = 1, 2, \ldots, n$ will represent subject (cluster, or mouth); and $j = 1, 2, \ldots, m_i$ will denote the site (e.g. tooth or tooth surface) within subject. Let Y_{ij} denote the response for site j in subject i; let $\underline{X}_{ij}$ denote the vector of predictor variables associated with fixed effects (vector β) for site j in subject i; and denote by $\underline{Z}_{ij}$ vector of the covariates associated with the random effects (vector $\underline{b}_i$ for subject i. The general model is given by:

$$h(\mu_{ij}) = \underline{X}_{ij}{}^T \underline{\beta} + \underline{Z}_{ij}{}^T \underline{b}_i \tag{13.2}$$

The random effects typically represent unobserved characteristics of the subjects. Many different random effects models could be assumed. In the simplest case, there is only one random effect: the random intercept, b_{0i}. In the school setting, the random intercept b_{0i} may represent the unobserved level of ability of the group of students in the i^{th} classroom when evaluating the performance of the students on a test. While one might measure certain aspects of the ability of the individual students, this can never be done perfectly; moreover, students from the same class share other characteristics (such as similar socioeconomic backgrounds, environmental or neighborhood exposures, diet, etc.) which are not measured perfectly (or perhaps not at all). In the oral health context, the cluster is the mouth and teeth are the sites in the cluster. For a study on the impact of dietary habits on caries experience, for example, the random intercept b_{0i} represents the i^{th} subject's susceptibility to caries, which might depend on genetic factors and environmental influences, whether these have been measured in the study or not.

The simplest random effects model is the random intercept model with one predictor:

$$h(\mu_{ij}) = b_{0i} Z_i + \beta_1 X_{ij} = b_{0i} + \beta_1 X_{ij} \tag{13.3}$$

where μ_{ij} represents the mean of y_{ij} given the random intercept b_{0i} (the condition 'given the b_{0i}' is important interpretationally; see below). Note that we further assume that $b_{0i} = \beta_0 + u_{0i}$, where the u_{0i} are distributed as normal with mean zero and variance τ^2, and where the Z_{ij} are equal to 1 for all i and j. The deviation of the observed response Y_{ij} from the predicted obtained by Equation (13.3) is called the residual error r_{ij} which is assumed to follow distribution G with mean zero. We also assume that the random effects are independent of the residual errors. Model (13.3) implies that there is correlation among the responses Y_{ij} from the same subject (keeping subscript i fixed). Indeed, for normally distributed random effects and residual errors (i.e.$(r_{ij} \sim N(0, \sigma^2))$ in a model for continuous responses, the correlation among responses within a cluster is given by: $\tau^2/(\tau^2 + \sigma^2)$ and is the intra-class correlation.

In longitudinal studies the random intercept model implies that the response of each subject deviates from the population average simply by the term u_{0i}, independent of time. Another popular longitudinal model is the random intercept and random slope model, where the regression coefficient for time (the only covariate) is also assumed to be random:

$$h(\mu_{ij}) = \beta_0 + \beta_1 t_{ij} + u_{0i} + u_{1i} t_{ij},$$

where t_{ij} is the time of the j^{th} observation made on subject i, and u_{1i} is the random slope that measures the i^{th} subject's deviation from the overall population average with respect to Y_{ij}'s linear dependence on time.

Example using the random effects approach to analyze continuous responses We can use the random intercept model to analyze the HIV Study data, specifying for each subject a random intercept, representing that subject's tendency to have a high or low IgM level over all the sites in his/her mouth. Two models will be considered. In both models the outcome is the natural logarithm of IgM level (the log transformation yields a distribution for outcome that is more closely normal), and predictors will include the subject's HIV status S (a subject-level variable) and the attachment level A (a site-level variable). The first random effects model is:

$$\text{MODEL 1} : h(\mu_{ij}) = b_{0i} + \beta_1 S_i + \beta_2 A_{ij}. \tag{13.4}$$

In the second model:

$$\text{MODEL 2} : h(\mu_{ij}) = b^*_{0i} + \beta^*_1 S_i + \beta^*_2 A_{ij} + \beta^*_3 \overline{A}_i \tag{13.5}$$

we add the mean whole-mouth attachment level $\overline{A}$ (a subject-level variable) as a predictor. The reason for including $\overline{A}$ is that adjusting for the average of site-specific exposure measurements can help us to distinguish subject-level effects from site-level effects; see Neuhaus and Kalbfleish (1998), Berlin *et al.* (1999), Raudenbush and Bryk (2002), and Begg and Parides (2003) for more details.

The data have been analyzed first using ordinary linear regression (with identity link and normal residual errors) ignoring the clustering, and then using random effects regression with a random intercept to account for correlation among sites within subject. Tables 13.1 and 13.2 show the output obtained using STATA software for this analysis. Table 13.1 gives the results from the naïve analysis (using ordinary linear regression) for Models 1 and 2. Table 13.2 gives the corresponding random intercept model results.

We can use the output in Tables 13.1 and 13.2 to characterize the relationships between immune response at the site (IgM), and the covariates. Focusing on HIV status, the naïve analysis for Model 1 tells us that the expected change in the log-transformed IgM level is 0.60, with a standard error (SE) of 0.10. Model 2 gives an estimate (SE) of 0.52 (.11). On the original scale for IgM, these equate to increases of about 82 % ($e^{0.60} \approx 1.82$) and 68 % ($e^{0.52} \approx 1.68$), indicating an augmented inflammatory response associated with HIV infection. The estimated

Table 13.1 Ordinary linear regression results for Models 1 and 2 applied to the HIV Study data (status = whether or not HIV infected, al = attachment level, mean_al = mean attachment level).

(a) Model 1

```
. reg log_igm status al
[partial output deleted]
------------------------------------------------------------------------------
     log_igm |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |   .6033568   .1048059     5.76   0.000      .397283    .8094306
          al |   .2837124   .0374417     7.58   0.000      .210093    .3573319
       _cons |   3.262343   .1493197    21.85   0.000     2.968744    3.555941
------------------------------------------------------------------------------
```

(b) Model 2

```
. reg log_igm status al mean_al
[partial output deleted]
------------------------------------------------------------------------------
     log_igm |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |    .523646   .1057633     4.95   0.000     .3156879    .7316041
          al |   .1678861   .0495586     3.39   0.001     .0704409    .2653312
     mean_al |   .2598879    .074235     3.50   0.001     .1139226    .4058531
       _cons |   2.772356   .2030821    13.65   0.000     2.373044    3.171668
------------------------------------------------------------------------------
```

standard errors, however, may be misleading due to the failure to account for clustering among sites. Table 13.2 gives roughly similar effect estimates, but the confidence limits are strikingly different; for Model 1, we get 0.70 (0.23) for the mean change in log IgM associated with HIV infection, and 0.58 (0.23) for Model 2. While the effect estimates are only slightly higher from the random effects model (12–17 % higher), the standard errors are dramatically increased (110 % higher), more than doubling in size in the random effects analysis. Consequently, the confidence intervals obtained from the random effects model are much wider than those from the naïve analysis. Throughout this analysis, the effect of HIV status remains significant after adjustment for intra-cluster correlation; but it is easy to see how adjustments for clustering can impact subsequent inferences in general.

A similar analysis of the site-specific attachment level measurement (AL) demonstrates that for predictor variables that vary from site to site within subject, the analysis that accounts for clustering can actually result in a smaller standard error for an effect estimate of interest. Based on the naïve analysis (Table 13.1), the parameter estimate (SE) for AL is 0.28 (0.04) from Model 1, and 0.17 (0.05) from Model 2. Both indicate a positive association between periodontal disease severity and immune response at the site. The estimate from Model 2 is somewhat smaller than that from Model 1, reflecting the effect of adjusting for whole-mouth attachment level, and providing a more accurate estimate of the within-subject effect (as opposed to the between-subject effect, represented by the regression coefficient for whole-mouth average attachment level). The random effects analysis presented in Table 13.2 reveals parameter estimates (SE) of 0.20 (0.04) and 0.17 (0.04) for

Table 13.2 Random intercept model results for Models 1 and 2 applied to the HIV Study data. (al = attachment level, mean al = mean attachment level).

(a) Model 1

```
. xtreg log_igm status al, i(id)
Random-effects GLS regression                   Number of obs      =       382
Group variable (i): id                          Number of groups   =        29

R-sq:  within  = 0.0441                         Obs per group: min =         9
       between = 0.4555                                        avg =      13.2
       overall = 0.2182                                        max =        16

Random effects u_i ~ Gaussian                   Wald chi2(2)       =     38.18
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
     log_igm |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |   .7011775   .2250997     3.11   0.002     .2599902    1.142365
          al |    .195138   .0391965     4.98   0.000     .1183143    .2719617
       _cons |   3.568102   .2177865    16.38   0.000     3.141248    3.994956
-------------+----------------------------------------------------------------
     sigma_u |  .54588091
     sigma_e |  .82250537
         rho |  .30578312   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

(b) Model 2

```
. xtreg log_igm status al mean_al,  i(id)
[partial output deleted]
------------------------------------------------------------------------------
     log_igm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |   .5822854   .2331706     2.50   0.013     .1252795    1.039291
          al |   .1678861   .0416904     4.03   0.000     .0861745    .2495977
     mean_al |   .2236211   .1194246     1.87   0.061    -.0104468    .4576889
       _cons |      2.893   .4208404     6.87   0.000     2.068168    3.717832
-------------+----------------------------------------------------------------
     sigma_u |  .54588091
     sigma_e |  .82250537
         rho |  .30578312   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Models 1 and 2, respectively. While we saw an increase in the size of the standard errors for HIV status (a subject-specific variable) going from the naïve to the adjusted analysis, the adjusted standard errors for attachment level (a site-specific variable) are about the same size or smaller than the standard errors obtained from the naïve analysis. This is likely because an analysis of a site-specific predictor variable can yield a gain in efficiency and precision compared to a subject-specific predictor, as comparisons can be made within subject, reducing heterogeneity from between-subject sources.

Example using the random effects approach to analyze binary responses Let's also investigate the IgM response as a binary variable (igmhi, equal to 1 if IgM is greater than or equal to 100 units and 0 otherwise). Just as we can specify a random intercept model with identity link and normal residual errors for continuous data, we can specify a random intercept model with logit link and binomial errors for binary data. In this model, the random intercept represents each subject's (cluster's)

Table 13.3 Ordinary and random intercept logistic regression model results for dichotomized IgM level as a function of HIV status, site-specific attachment level, and whole-mouth mean attachment level as covariates (Model 2).

(a) Logistic model for High IgM, ignoring intra-cluster correlation

```
. logit igmhi status al mean_al

[partial output deleted]
```

igmhi	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
status	1.106018	.2394114	4.62	0.000	.6367804	1.575256
al	.3225491	.1250596	2.58	0.010	.0774367	.5676615
mean_al	.8477136	.2306247	3.68	0.000	.3956974	1.29973
_cons	−5.066748	.7614735	−6.65	0.000	−6.559209	−3.574288

(b) Random intercept logistic model for high IgM

```
. xtlogit igmhi status al mean_al, i(id)

[partial output deleted]
```

igmhi	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
status	1.768829	.662967	2.67	0.008	.4694375	3.06822
al	.4153407	.1435334	2.89	0.004	.1340203	.696661
mean_al	.8679132	.4119206	2.11	0.035	.0605636	1.675263
_cons	−6.066521	1.517745	−4.00	0.000	−9.041246	−3.091796
/lnsig2u	.6911215	.4506593			−.1921546	1.574398
sigma_u	1.412782	.3183417			.9083938	2.197233
rho	.3776051	.1059137			.2005273	.5947291

```
Likelihood-ratio test of rho=0: chibar2(01) =    43.55 Prob >= chibar2 = 0
```

likelihood of a positive response. While the residual errors are assumed to be binomial, the random intercept logistic regression model typically assumes that the intercepts b_{0i} follow a normal distribution with mean β_0 and variance τ^2, the same assumption made in the linear model for continuous responses.

Table 13.3a gives the results of an ordinary logistic analysis of high IgM, a binary outcome, as a function of the three above described covariates. This model reveals an estimated odds ratio of approximately 3.02 ($= e^{1.1060}$) for HIV status, indicating that HIV infection increases the odds of a strong immune response at a site by a factor of 3. For attachment level, we find an estimated odds ratio of 1.38, representing a 38 % increase in the odds of a heightened immune response at the site for every one-millimeter increase in attachment level. But this analysis ignores clustering among observations within a subject;, therefore the standard errors (and confidence limits and test statistics) are not valid. To address this concern, we next fit a logistic model with random intercept term to the HIV study data. Doing so, we find a very strong effect estimate for status (see Table 13.3b). The log odds ratio estimate is 1.77, translating to an odds ratio estimate of 5.86, indicating that the odds of high IgM are more than 5 times higher with HIV infection. The 95 % confidence interval for this parameter ranges from 1.60 to 21.50. Looking

at attachment level, we see that an increase of one millimeter is associated with an odds ratio of $1.51 = \exp(0.4153)$, indicating a 51 % increase in the odds of elevated IgM with a one-millimeter increase in attachment level.

13.2.3.2 Final comments on the random effects approach

Note that the effect estimates for both status and attachment level appear stronger (further from the null) in the random effects logistic regression analysis than in the ordinary logistic analysis. This may be due to the difference in interpretation between an ordinary analysis for independent outcomes and a random effects analysis for clustered outcomes (Davis, 2002). In the ordinary analysis, regression parameters have a so-called 'population-averaged' interpretation, meaning that they represent the effect of the covariate(s) on the population mean response. This is the traditional interpretation for most analyses of non-clustered (independent) data, as well as the interpretation for parameters estimated via a GEE approach (see next section). In contrast, the regression coefficients from the random effects analysis have a 'subject-specific' interpretation, representing the effect of the covariate(s) on the individual subject's response (not the mean population response), over and above that captured by the random effect. (As noted by Fitzmaurice *et al*. (2004), the exposure being evaluated must take on different values within a subject for the latter interpretation to apply; only in this case will the distinct random intercept for each subject 'cancel out,' allowing for the above interpretation of the random effects regression parameter.) In general, random effects models yield regression coefficients that are greater (in absolute value) than the corresponding population averaged coefficients. The practical implication is that the regression coefficients of the two approaches may not be comparable in general, except in the case of the identity link where the two regression coefficients are often quite similar for large sample sizes.

13.2.3.3 The GEE approach

The GEE marginal regression approach, first proposed by Liang and Zeger (1986) and Zeger and Liang (1986), involves fitting a GLM-type model specifying a particular form for the correlation among observations within a cluster. The correlation parameters are estimated separately from the regression parameters, however, and are based on the residuals from the fitted model, with the estimation procedure iterating between the regression model estimation and correlation parameter estimation. As a final step, the observed correlation/covariance structure is compared to the assumed model-based structure; a so-called sandwich-type variance correction (Huber, 1967; White, (1980, 1982)) is then applied in order to further adjust the covariance terms and standard errors to best account for intra-cluster correlation. This correction multiplies the model-based covariance structure (the 'bread' of the sandwich) by the cross-product matrix of residuals (the 'meat' in the centre of the sandwich), thereby adjusting the covariance terms either by a smaller amount (when the model-based covariance structure is close to the empirical covariance structure)

or a larger amount (when the model-based and empirical covariance structures are further apart).

The theory behind this approach states that if one proposes the correct model form (i.e., the appropriate mean-covariate relationship), then consistent estimates for the regression coefficients will result. Furthermore, these parameter estimates are asymptotically normal and we can obtain estimates of the covariate terms. Finally, these results hold even if the proposed correlation structure is inaccurate, as the variance correction at the last step assures the validity of the standard error estimates (though there may be some loss in efficiency).

To fit a GEE model, three elements must be specified. First is the mean-covariate relationship, as captured by the link function: $h(\mu_{ij}) = {X_{ij}}^T \beta$. Second is the mean-variance relationship (or the distributional family of the residual error terms); for example, for the binomial, we would say that the variance of outcome Y_{ij} takes the form $\phi\mu_{ij}(1 - \mu_{ij})$, and for the normal, we assume $\text{Var}(Y_{ij}) = \phi$. Finally, we specify a 'working' or 'presumed' correlation structure for responses within a cluster. Remember that a properly specified GEE model will give valid results even when the correlation structure has been mis-specified; this is a result of the sandwich variance correction described earlier. Thus, a data analyst may specify any working correlation structure, confident that the accuracy of his selection will not threaten the validity of the analysis, given that the rest of the model has been specified appropriately. Common choices for working correlation include the independence correlation structure and the exchangeable correlation structure. The independence structure assumes zero correlation among items within a cluster; although this is clearly incorrect, the last-step, sandwich variance adjustment will correct for the inaccuracy. The exchangeable correlation structure assumes that the correlation between any two items from a single cluster (subject) have the same level of correlation. Although this may not seem like a natural choice for oral health studies in which correlation between sites may be supposed to decrease as distance between sites increases, researchers have shown that the exchangeable correlation structure is often 'close enough' to achieve valid and efficient results in most periodontal disease studies (Ten Have *et al*., 1995). Yet another choice is an auto-regressive correlation structure, which allows correlation to decline as a function of some metric (like distance between sites in the mouth).

Example using GEE approach to analyze continuous responses Let us use the GEE method to analyze the HIV Study data. As before, we will specify the log-transformed IgM measurement as the outcome in a linear model, with HIV status, attachment level at the site, and mean whole-mouth attachment level as predictors. For simplicity, we will employ all three predictors as covariates (as in Model 2), and compare the results obtained with GEE to the naïve results and those obtained using the random effects approach.

The STATA output is given in Table 13.4. There we see that the estimated effect and SE (in parentheses) for the HIV status variable is 0.58 (0.22) when we specify an exchangeable correlation. This is nearly identical to the estimate and standard error obtained via the random intercept model. Looking at a variable which changes

Table 13.4 GEE regression model results for log IgM level as a function of HIV status, site-specific attachment level, and whole-mouth mean attachment level as covariates.

```
. xtgee log_igm status al mean_al, i(id) family(gaussian) link(identity)
corr(exch) robust

Iteration 1: tolerance = .03817687
Iteration 2: tolerance = .00016253
Iteration 3: tolerance = 8.454e-07

GEE population-averaged model                   Number of obs      =       382
Group variable:                         id      Number of groups   =        29
Link:                             identity      Obs per group: min =         9
Family:                           Gaussian                     avg =      13.2
Correlation:                  exchangeable                     max =        16
                                                Wald chi2(3)       =     58.91
Scale parameter:                  .9481633      Prob > chi2        =    0.0000

                                  (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semi-robust
     log_igm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |   .5820725   .2245741     2.59   0.010     .1419153     1.02223
          al |   .1678861   .0622278     2.70   0.007     .0459218    .2898504
     mean_al |   .2237537   .1213156     1.84   0.065    -.0140205    .4615279
       _cons |   2.892563   .3671758     7.88   0.000     2.172911    3.612214
------------------------------------------------------------------------------
```

from site-to-site, we see that the parameter estimate (SE) for attachment level is 0.17 (0.06); again, this is almost the same as what we obtained using random effects regression. When we fit the model using an independence correlation structure, we obtain very similar results in this particular example. In sum, we find from the GEE analysis that being HIV positive is associated with elevated IgM at the site, adjusted for attachment level and whole-mouth mean attachment level. The estimated regression coefficient is 0.58, with 95 % confidence interval (0.14, 1.02) (from the model with exchangeable correlation). These estimates and confidence limits correspond to an increase on the IgM scale of 79 % with 95 % confidence interval (15 %, 177 %). In addition, an increase of one millimeter in attachment level is associated with 0.17 unit-increase in log IgM (95 % confidence interval from 0.05 to 0.29). This equates to about a 19 % increase in IgM, with 95 % confidence interval (5 %, 34 %).

Example using GEE approach to analyze binary responses One of the advantages of the GEE approach is that the same construct can easily accommodate a variety of response variables, including continuous responses, discrete responses, counts, and rates. To illustrate this, we re-analyze the HIV Study data using the dichotomized IgM (as <100 units versus ≥ 100 units) as the response variable.

Studying the output in Table 13.5, which adjusts for intra-subject correlation, we see that being HIV positive is associated with a higher risk of elevated IgM at the site, adjusted for attachment level at the site and whole-mouth mean attachment

Table 13.5 GEE logistic model for binary IgM, with exchangeable correlation and robust covariance.

```
. xtgee igmhi status al mean_al, i(id) family(binomial) link(logit) corr(exch)
robust

 [partial output deleted]

GEE population-averaged model                   Number of obs      =       382
Group variable:                         id      Number of groups   =        29
Link:                                logit      Obs per group: min =         9
Family:                           binomial                     avg =      13.2
Correlation:                  exchangeable                     max =        16
                                                Wald chi2(3)       =     17.41
Scale parameter:                         1      Prob > chi2        =    0.0006
                                 (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semi-robust
      igmhi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |   1.132119   .4484048     2.52   0.012     .2532614    2.010976
          al |   .3148698   .1337441     2.35   0.019     .0527362    .5770035
     mean_al |   .6515761   .4204612     1.55   0.121    -.1725127    1.475665
       _cons |  -4.292442   1.531155    -2.80   0.005     -7.29345   -1.291434
------------------------------------------------------------------------------
```

level. The estimated regression coefficient for the log odds ratio for status is 1.13, with 95 % confidence interval (0.25, 2.01). Exponentiating these values, we find that the estimated HIV status/IgM odds ratio is 3.10, with 95 % confidence limits from 1.29 to 7.47. Results were almost identical for both of the working correlation structures selected (exchangeable or independence). Note that the regression parameters from a GEE model have a 'population-averaged' interpretation, like those from an ordinary analysis of independent data, in that they represent the expected change in outcome for a one-unit change in the predictor variable across different individuals with different values of the predictor (or, according to Davis (2002), the expected effect of the explanatory variables on the population average). This is in contrast to the random effects approach, which produces 'subject-specific' estimates (where the regression parameters represent the 'within-individual' differences, above and beyond that captured by the random effects). Thus, we would expect the parameter estimates from the ordinary logistic and GEE logistic analyses to be fairly similar, and this is confirmed when we compare the output from Table 13.5 with Table 13.3a. While the effect estimates for the predictors that vary within subject (al and mean_al) are similar across all three modeling approaches, both the ordinary logistic and GEE logistic parameter estimates for HIV status (fixed within subject) are much lower than the corresponding estimate from the random effects analysis in Table 13.3b. This can occur when comparing marginal model output to random effects model output for non-linear regressions. Recall that we did not observe large differences between the GEE and random effects estimates for the linear model; this is because for the linear model only, the interpretation of GEE and random effects parameters is identical. Both can be viewed as subject-specific analyses. This is not the case for non-linear regression analysis, as demonstrated in the current example.

Final comments on GEE The GEE approach is useful for analyzing clustered response data of many different types (continuous, binary, counts, rates, etc.). It is more flexible than the random effects approach, insofar as the basic model structure and mechanics are the same, regardless of the type of response variable or assumed error variance; while there are random effects model generalizations for non-continuous data, these require, in general, different structures or assumptions compared to the model for continuous responses. GEE regression results are valid whether or not the working correlation structure has been correctly specified (assuming that the basic model formulation, with respect to mean-covariate relationships, is correct). While validity is preserved when the wrong working correlation is specified, efficiency may be lost. Efficiency is improved as the working correlation structure gets closer to the true underlying correlation structure.

13.2.4 Bootstrap resampling for clustered data

Bootstrapping is 'a computer-based method for assigning measures of accuracy to statistical estimates' (Efron & Tibshirani, 1993). The basic idea behind bootstrapping is to sample with replacement over and over again from the original dataset, creating many new 'bootstrap' samples of the same size as the original, and then to draw inferences about the parameter(s) of interest based on the distribution of the observed parameter estimates from the multiple bootstrap samples. The number of bootstrap samples, say B, should be large, and is typically in the range of 1000–2000 (Efron & Tibshirani 1993). In this way, the B sample estimates of the parameter of interest can be computed, and their distribution studied in order to derive estimates of the variability of the sample estimator. This information can be used to determine the standard error or a confidence interval for the desired parameter, whether it is the mean, correlation, proportion, odds ratio, or regression coefficient. Statisticians rely on bootstrapping when the parametric assumptions behind a particular procedure are in doubt, or when the situation calls for more complicated calculations for the standard errors. The latter case is the most important issue in the analysis of clustered data, where obtaining valid standard error estimates is the biggest inferential challenge.

To obtain standard errors that appropriately account for intra-cluster correlation in the clustered data setting, we can use the bootstrap approach, so long as we resample *by cluster* rather than by observation. Thus, in bootstrapping with clustered data, we can resample the clusters, taking all the members of the selected cluster, always selecting the same number of clusters in each bootstrap sample. In this way, we mimic the sampling structure of the original dataset, and preserve the validity of resulting inferences. This approach was originally proposed in the complex survey literature (e.g. Gross, 1980; McCarthy & Snowden, 1985) and in the time series literature (Kunsch, 1989); and has been further explored for correlated data in more general settings by Feng, McLerran and Grizzle (1996) and by Field and Welsh (2007). Note that in our example, we apply the simplest method of resampling by cluster, as described above. (Note that other approaches

Table 13.6 Bootstrap regression analysis results for continuous IgM level as a function of HIV status, site-specific attachment level, and whole-mouth mean attachment level as covariates (Model 2).

```
. bootstrap "reg log_igm status al mean_al" _b, reps(1000) dots cluster(id)

command:      reg log_igm status al mean_al
statistics:   b_status   = _b[status]
              b_al       = _b[al]
              b_mean_al  = _b[mean_al]
              b_cons     = _b[_cons]
..................................................................................
..................................................................................
[and so on…]

Bootstrap statistics                                Number of obs    =       382
                                                    N of clusters    =        29
                                                    Replications     =      1000
------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
    b_status |  1000   .523646  .0161417  .2336487   .0651475   .9821445   (N)
             |                                      .0761958   .9968094   (P)
             |                                      .0643689   .9492046  (BC)
        b_al |  1000  .1678861 -.0037638  .0620358   .0461506   .2896216   (N)
             |                                      .0444895   .2807251   (P)
             |                                       .051232   .2856158  (BC)
  b_mean_al  |  1000  .2598878  .0209781  .1533635  -.0410637   .5608394   (N)
             |                                      .0276329   .6699888   (P)
             |                                      .0251417   .6556358  (BC)
      b_cons |  1000  2.772356 -.0615937  .4951104    1.80078   3.743931   (N)
             |                                      1.431676   3.457677   (P)
             |                                      1.403653   3.455079  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

are also possible; see Feng *et al*. (1996) for more details.) So if clusters are of different sizes, then the actual *number of observations* in the bootstrap samples will vary somewhat; this can become problematic if the number per cluster varies widely. Feng and colleagues have shown that this method typically performs well regardless of cluster size once the number of clusters reaches 50 or more.

The bootstrap approach can be applied to any statistical procedure. For our purposes, we are most interested in using the bootstrap to obtain regression parameter estimates. This is accomplished by selecting a large number of clustered bootstrap samples, fitting a regression model to each sample, then examining the distribution of regression parameter estimates. In this way, we can look at the mean regression parameter estimates over all samples, and also obtain the variance estimates for the regression parameters in question, and generate confidence intervals for the regression parameters based on percentiles from the bootstrap distribution.

13.2.4.1 Bootstrap approach applied to the HIV study data

Table 13.6 gives the output from the bootstrap analysis of continuous IgM using linear regression. We also ran a bootstrap analysis of binary IgM by running a

logistic regression bootstrap (output not shown). Note that in both cases, we give the STATA command for the 'ordinary' (unclustered) regression model. This is because the clustering will be accounted for in the sampling distributions of the resulting parameter estimates, so long as we resample by cluster, and not by individual observation. The estimated regression coefficient for HIV status from the first analysis is about 0.52, with a percentile-based 95 % confidence interval of (0.08, 1.00). Corresponding estimates and confidence intervals from the random effects approach and the GEE (exchangeable correlation) approach are 0.58 (0.13, 1.04) and 0.58 (0.14, 1.02), respectively. The bootstrap confidence interval is about the same width as those obtained via GEE and random effects, but slightly lower in magnitude, consistent with the parameter estimate. The bootstrap estimate of the log odds ratio from the logistic analysis (data not shown) is 1.11, with 95 % confidence interval (0.24, 2.31). From the GEE analysis with exchangeable correlation, we previously obtained 1.13 and (0.25, 2.01). The bootstrap confidence interval is somewhat wider than the GEE confidence interval, but effect estimates are similar; both GEE and bootstrap give an HIV status effect that is more modest than that obtained via RE modeling.

Overall, we observe a significant association between HIV status and IgM, adjusted for site-specific attachment level and mean whole-mouth attachment level, whether IgM is analyzed as a continuous or binary outcome variable. We obtained results via bootstrap that are consistent with those obtained via random effects and GEE analysis. Conducting bootstrap analysis in STATA proves to be straightforward, and constitutes a viable alternative to generalized linear models for clustered data without having to rely on complicated and unverifiable assumptions regarding, for example, independence of observations, normality of sample regression coefficients, or underlying correlation structure. A disadvantage of the bootstrap approach is that it is based on random sampling; and, as such, the results from separate runs of the 'bootstrap' command can result in slightly different summary statistics for the parameters of interest. Concern over variability might be addressed by increasing the number of bootstrap samples.

13.3 Final remarks and further reading

Clustered data offer the opportunity to thoroughly study within-subject and between-subject effects in oral health research. While clustering presents certain challenges in the analysis (primarily that of obtaining valid standard error estimates, confidence intervals, and test statistics in the presence of intracluster correlation), use of appropriate methods can lead us to a better understanding of the complex factors and interrelationships at play, with the ultimate benefit of drawing more reliable and robust conclusions regarding disease risk factors, determinants of progression, and identification of effective prevention and treatment strategies. We have presented a number of methodological approaches and STATA software code for analyzing clustered data. Below we make a few more notes regarding their use and interpretation.

This chapter has focused squarely on the analysis of clustered data from oral health studies. Under oral health study designs, researchers often collect information on multiple sites within a single subject, introducing correlation, generally positive, among responses within subject. We have devoted much less attention to the issue of longitudinal studies, in which multiple observations of the same variable are recorded repeatedly over time in the same subject. While this application is somewhat different from the oral health scenario, the vast majority of concepts and approaches apply equally well in either setting. Clearly the same types of statistical methods and interpretation of results can be utilized in oral health and longitudinal studies in order to make valid inferences about the relationships between the response measurements and the explanatory variables.

Always keep in mind that missing data can introduce bias into an analysis of clustered data, just as missing data can bias any other type of analysis, clustered or not. If missingness is non-ignorable, then resulting effect estimates may be misleading. Researchers should be encouraged to evaluate missingness to the extent possible in a given data set. With clustered data, there may be more information to gauge the representativeness of the missing observations than is typical in a nonclustered analysis of independent data. For example, in a study that records data from multiple sites within the mouth, we might look to see whether sites surrounding the missing point differ in mean response levels compared to sites further away. In a longitudinal study, we can compare early response measurements from subjects who drop out to those from subjects who remain on study to identify any systematic differences. In either setting, we can evaluate whether covariate values at the site are related to the likelihood of a missing response using, for example, logistic regression with the outcome defined as 'missing' versus not. Further, the careful data analyst might conduct a small 'sensitivity analysis,' imputing either very good or very poor responses at the missing sites and refitting regression models to assess the degree to which the regression coefficients are affected. Taking these steps will help to guide the researcher in evaluating the impact of missing data on the overall study conclusions.

Finally, the presentation in this chapter is fairly brief; for a more comprehensive look at techniques for analyzing clustered data, you may consult a number of excellent texts on the subject that have been published over the past fifteen years, including those by Fitzmaurice *et al*. (2004), Diggle *et al*. (2002), Davis (2002), Davidian and Giltinan (1995), Singer and Willett (2003), and Verbeke and Molenberghs (1997).

References

Begg, M.D. & Parides, M.K. (2003) Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine* **22**: 2591–2602.

Begg, M.D. & Paykin, A.B. (2001) Evaluation of performance and software for a modified Mantel-Haenszel statistic for correlated data. *Journal of Statistical Computation and Simulation* **70**: 175–95.

Berlin, J.A., Kimmel, S.E., Ten Have, T.R., & Sammel, M.D. (1999) An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty. *Biometrics* **55**: 470–6.

Davidian, M. & Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall/CRC.

Davis, C.S. (2002) *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer-Verlag.

Diggle, P.J., Heagerty, P., Liang, K.Y., & Zeger, S.L. (2002) *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Efron, B. & Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.

Feng, Z., McLerran, D., & Grizzle, J. (1996) A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine* **15**: 1793–1806.

Field, C.A. & Welsh, A.H. (2007) Bootstrapping clustered data. *Journal of the Royal Statistical Society, Series B* **69**: 369–90.

Fitzmaurice, G.M., Laird, N.M., & Ware, J.H. (2004) *Applied Longitudinal Analysis*. New York: John Wiley & Sons, Inc.

Fleiss, J.L., Park, H., & Chilton, N.W. (1987) Within-mouth correlations and reliabilities for probing depth and attachment level. *Journal of Periodontology* **58**: 460–63.

Fleiss, J.L., Wallenstein, S., Chilton, N.W., & Goodson, J.M. (1988) A re-examination of within-mouth correlations of attachment level and of change in attachment level. *Journal of Clinical Periodontology* **15**: 411–14.

Gross, S. (1980) Median estimation in sample surveys. *Proceedings of the Survey Methods Section of the American Statistical Association.* American Statistical Association, pp. 181–4.

Huber, P. (1967) The behaviour of maximum likelihood estimators under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*. University of California Press.

Kunsch, H. (1989) The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17**: 1217–41.

Lamster, I.B., Holmes, L.G., Gross, K.B., *et al*. (1994) The relationship of beta-glucuronidase activity in crevicular fluid to clinical parameters of periodontal disease. Findings from a multicenter study. *Journal of Clinical Periodontology* **21**: 118–27.

Laster, L.L. (1992) Some aspects of efficient experimental design and analysis in periodontal trials. *Journal of Periodontal Research* **27**: 405–11.

Liang, K.Y. & Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrics* **73**: 13–22.

McCarthy, P.J. & Snowden, C.B. (1985) The bootstrap and finite population sampling. *Vital and Health Statistics*, 2-95, Washington, DC: US Government Printing Office.

Neuhaus, J.M. & Kalbfleisch, J.D. (1998) Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54**: 638–45.

Raudenbush, S.W. & Bryk, A.S. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edn. Sage Publications, Inc.

Singer, J.D. and Willett, J.B. (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press.

Ten Have, T.R., Landis, J.R., & Weaver, S.L. (1995) Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Statistics in Medicine* **14**: 413–29.

Verbeke, G. & Molenberghs, G. (1997) *Linear Mixed Models for Longitudinal Data*. New York: Springer.

White, H. (1980) A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**: 817–50.

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* **50**: 1–25.

Zeger, S. & Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**: 121–30.

14

Missing data and informative cluster sizes

Stuart A. Gansky and John M. Neuhaus

14.1 The problem of missing data and dropouts

Often the largest source of error and variability in studies is not from the observed information, but rather from important, influential information which for several reasons was unobtainable. Such is the case with incomplete or missing data which not only have obvious implications for reduced power (efficiency) through smaller sample size, but also, perhaps more importantly, can introduce substantial parameter estimate bias (Little & Rubin, 1987). For example, in a longitudinal study, people in one group may be much less likely than another group to return for visits (lost-to-follow-up) because their oral health is good; conversely, in other studies, those with the worst health may find care elsewhere or be homebound. In either scenario, missing data can greatly influence standard change scores that are computed. Fortunately, various methods to account for missing data have been developed and, through modern computing, practical analyses for handling these situations are possible.

In the last twenty years statisticians have developed many methods to deal with missing data providing detailed insights on missing data's impact on statistical analyses. Since much of these statistical developments is quite technical in nature, many assertions in this chapter have deep statistical roots which cannot be explained in detail here.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

14.1.1 What is it, why is it important, and what can be done?

Missing data is perhaps inevitable in real-life studies whether cross-sectional or longitudinal. Table 14.1 presents a fictitious example with 5 variables. If all 5 were measured at the same time, the example would be cross-sectional. If the first 3 variables were measured at baseline and the last 2 being follow-up measures of the third variable, the example would be longitudinal. In cross-sectional or longitudinal studies, specific data items might be missing due to study staff problems (e.g. lost or damaged biological samples such as saliva or plaque), participants skipping questionnaire items (e.g. income or citizen status), participants truncating the visit (e.g. an uncomfortable periodontal probing), participants opting out of a procedure (e.g. venipuncture), or participants with a health condition contraindicating a specific study procedure (e.g. artificial heart valve recipient and periodontal probing). If the reason data are missing does not relate to the health condition being measured (e.g. a sample lost by staff), the ramifications are not as serious as when the reason relates to the underlying health or a factor related to health (e.g. refusing to report income because it is very low or very high). However, even a small amount of missing data on many items can reduce sample size greatly in multivariable models. The process or way by which data become missing is called the missingness mechanism.

It is useful to distinguish among several missing data patterns. If missing values are from subjects ending a cross-sectional interview before completion, those incomplete items are called *monotone missing*. In longitudinal studies, entire visits might be missing either intermittently or from participants dropping out altogether due to scheduling difficulties, relocating, withdrawing consent, worsening health, or unknown reasons. When a participant misses a visit and all subsequent visits, such as subjects 4 and 5 in Table 14.1, the person is a *dropout* for the remainder of the study, which is the simplest example of monotone missing data. In general, if p variables for the n participants can be ordered by the amount of missing data from smallest to largest $(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p)$ and all participants who are missing one variable (e.g. $\boldsymbol{y}_2$) are also missing all variables with more missingness (e.g. $\boldsymbol{y}_3, \ldots, \boldsymbol{y}_p$), then it is considered monotone missing as illustrated in the schematic of Figure 14.1. The variables in Table 14.1 would be monotone missing (by moving y_{3i} before y_{1i}) without the last subject, who in the longitudinal scenario missed the first follow-up but not the second. If it is *intermittent missing* and cannot be rearranged to be monotone, it is denoted as *arbitrary* or *non-monotone missing*, as is the data in Table 14.1. In general, it is easier to statistically correct for monotone missing values than for arbitrary ones and more analytic approaches exist for the former than the latter.

Missing data can produce several problems for statistical analysis. First, missing data results in loss of *power* (efficiency) of statistical tests. Often researchers will increase planned sample size to allow for missing data (e.g. dropout). However, an issue of greater concern is that missing data can introduce bias. For example, people with especially low or high socioeconomic status are often more reluctant to report income, which relates to many health conditions. Moreover, people with worse

Table 14.1 Sample subset for 5 variables ($y_{1i} - y_{5i}$) with missing data (•) and missing data indicators (HS = high school; K = 1000).

Subject	Variables					Missing indicators				
	y_{1i}	y_{2i}	y_{3i}	y_{4i}	y_{5i}	r_{1i}	r_{2i}	r_{3i}	r_{4i}	r_{5i}
1	≥HS	>$50K	0	0	1	0	0	0	0	0
2	<HS	<$20K	3	9	•	0	0	0	0	1
3	≥HS	$20-50K	0	0	•	0	0	0	0	1
4	<HS	•	5	•	•	0	1	0	1	1
5	•	•	2	•	•	1	1	0	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	≥HS	$20-50K	1	•	5	0	0	0	1	0

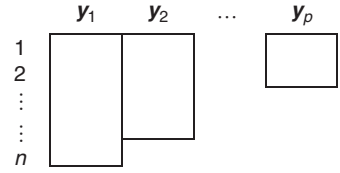


Figure 14.1 Schematic of a monotone missing data pattern.

disease may return less often for visits to receive study treatment (as illustrated in the last 3 variables of Table 14.1) or those with improving health may be more compliant in general and more likely to return for visits (see the example in Section 14.1.4.2). Thus, data that are missing often contain some important aspect that the non-missing data do not have. Fortunately, there are a variety of analytic methods for handling missing data that either allow estimating this important information or at least allow understanding its potential impact.

14.1.2 Prevention and planning

The best way to handle missing data is to design and perform studies to limit missing data as much as possible (DeLeeuw, 2001). Various approaches include recruiting participants who would be very likely to complete study procedures, not overburdening study participants with unessential components of the study, providing incentives for participants, and developing a rapport with study participants. However, care must be taken not to exclude too many participants to preserve generalizability to the target population (see Chapters 5–7).

14.1.3 Analytic remedies

Several approaches are available, including complete case (CC) analysis, imputation techniques, likelihood-based analyses (such as linear and generalized linear mixed

models), weighted analyses, and sensitivity analyses using nonignorable models. These approaches are described in this chapter along with illustrative examples.

14.1.4 Examples

As with most real-life research, missing data commonly arise in oral health studies including clinical trials, surveillance surveys and observational epidemiologic studies, as the following examples indicate.

14.1.4.1 Fluoride varnish early childhood caries prevention trial

A recent randomized parallel groups controlled trial of fluoride varnish (FV) to prevent early childhood caries (Weintraub *et al*., 2006) randomly assigned 376 children, ages 6–44 months and caries-free at baseline, into one of three arms stratifying on the two clinics: counseling only (0 FV control), counseling plus 2 FV applications (i.e. every 12 months), and counseling plus 4 FV applications (i.e. every 6 months). A single pediatric dentist blinded to treatment group performed full mouth 1- and 2-year follow-up dental examinations. A total of 183 (48 %) children had both follow-up exams, 78 (21 %) had only a 1-year follow-up, 19 (5 %) had only a 2-year follow-up, and 96 (26 %) had neither; thus, the missingness pattern is nonmonotone (arbitrary). Children missing a follow-up tended to have worse dental health than those with complete follow-ups. For ethical reasons, children with incident caries at the 1-year follow-up were given therapeutic FV treatment and exited from the prevention study, which (as defined in Section 14.2.2) is a missing at random (MAR) missing data mechanism (see below) for the dropouts since missing the 2-year follow-up depended on the 1-year follow-up observations.

14.1.4.2 California Oral Health Needs Assessment of Children (COHNAC, 2004-5) complex sample survey

COHNAC (also known as the California Smile Survey) was a complex health examination survey ($n = 21{,}399$) in 21 regional strata and 186 school clusters in California in 2004–5 (Dental Health Foundation, 2006). Survey sampling weights were based on inverse probability of selection of kindergarteners and third graders as well as non-response corrections. Since passive consent (right of refusal) was used in 80 % of schools, only 47 % of parents returned questionnaires, so potentially important covariates were missing for 53 % of children. However, school-level information as well as examiner recorded child-level information was available for most children.

14.1.4.3 Intergenerational epidemiologic cohort study of adult periodontitis

An observational epidemiologic cohort study of periodontal (gum) disease performed full mouth dental examinations on participants (Gansky *et al*., 1999). Participants had periodontal attachment level (millimeters) for 6 sites around posterior teeth present (first and second premolars and molars); mean attachment level and

worst loss of attachment (WLA) per tooth were calculated as was dichotomized tooth-level WLA $\geq$3 mm, indicating mild periodontal disease. Third molars (wisdom teeth) were excluded since they are seldom extracted for disease. The 406 participants with 2 or more posterior teeth periodontally probed in these analyses had cluster sizes (numbers of teeth) from 2 to 16 (mean 12, median 13). Missing within-cluster observations (i.e. smaller cluster size) appeared related to disease levels: people with fewer teeth tended to have worse attachment (Hoffman *et al*., 2001; Williamson *et al*., 2003; Neuhaus & McCulloch, 2006).

14.2 Missing data terminology and patterns

The standard missing data terminology of Little and Rubin (1987) will be used in this chapter to describe the missingness mechanisms (the process by which the data became missing). To describe these mechanisms, we first partition the intended response for the ith subject, y_i, into observed, $y_{i(obs)}$, and missing, $y_{i(mis)}$, components where $y_{i(obs)}$ is the observed value for the ith subject and $y_{i(mis)}$ is the missing value for the ith subject. (Note that in practice $y_{i(mis)}$ will never be observed.) Additionally, we define a binary random variable, r_i, to indicate if y_i was observed ($r_i = 1$) or not ($r_i = 0$), as in Table 14.1. The *missingness mechanism* is defined as the probability of being missing given the observed and missing data: $P(r_i|y_{i(obs)}, y_{i(mis)})$. The missing value mechanisms can be described by their relationship to observed and missing responses. Note that all these methods assume that models correctly specify covariates. (e.g. in a regression model all necessary covariates are included in the right way. Since in practice this often is not the case, actual analyses may be even more complex than discussed here.)

14.2.1 Missing Completely At Random

The term *Missing Completely At Random* (MCAR) means the probability of missing is independent of all responses – both those (previously) observed and those unobserved (missing): $P(r_i|y_{i(obs)}, y_{i(mis)}) = P(r_i)$. MCAR means missing data are a representative sample of the full data (observed and unobserved). Under the MCAR assumption, a simple mean of the observed data, $y_{i(obs)}$, is an unbiased estimate of the full data. This is the most restrictive scenario, but the assumption made by many common statistical methods (e.g. t-test, analysis of variance, Wilcoxon rank sum, and generalized estimating equations (GEE)). The MCAR assumption can be tested with real data. An example of MCAR in oral health research is the dental examination protocol in the National Health and Nutrition and Examination Survey (NHANES) which randomly selects one side of the maxilllary arch and one side of the mandibular arch (quadrant) for periodontal examinations (i.e. two of the four quadrants, stratifying on arch). Another example might be a randomized clinical trial with a delayed start which is unable to perform the final visit for, say, the last 10 % of participants. In both cases, the missing or unobserved data relate neither to what those values would have been nor to the prior observations.

It is possible to test whether the missingness mechanism is MCAR against a MAR process, explained in the next section.

14.2.2 Missing At Random

Missing At Random (MAR) is defined as missingness only relating to observed values (not relating to the actual missing values after conditioning on the observed values). Thus, MAR means $P(r_i|y_{i(obs)}, y_{i(mis)}) = P(r_i|y_{i(obs)})$. An example of a MAR mechanism in a RCT occurs when participants drop out of the study from lack of therapeutic effect, since earlier observed value(s) relate to the actual health status (the value the unobserved missing measure would have been). Another example is a study with multiple-stage screening: the data missing for the subsequent stages are MAR since they are missing due to observed values for the initial stages i.e. initial negative screening test). An important question is whether and when the missing data mechanism distorts parameter estimation. For instance, when an ordinary regression model uses data with missing observations, a vital question is if the regression coefficient estimates are biased from the missing data mechanism. If not, then the missing data mechanism is called *ignorable*. Both MCAR and MAR are *ignorable* if the parameters of the statistical model for the missingness mechanism, $P(r_i|y_{i(obs)}, y_{i(mis)})$, are independent of the parameters of interest and the parameters are estimated with direct likelihood. For instance, suppose that $P(r_i|y_{i(obs)}, y_{i(mis)})$ is given by a logistic regression model with regression coefficients in vector $\boldsymbol{\alpha}$ (i.e. the probability of missingness depends on certain covariates) and that the model of interest relating y_i to covariates is an ordinary regression model with regression coefficients in vector $\boldsymbol{\beta}$. Then ignorability occurs for a MAR model when all $\boldsymbol{\alpha}$-regression coefficients are independent of all $\boldsymbol{\beta}$-regression coefficients. In that case likelihood-based procedures (e.g. generalized linear mixed effects models; see Section 14.3.3 and Chapter 13) can generate consistent estimates (estimates close to true values for large sample sizes). The reason is that the total likelihood (of measurement process and of missing data process) splits up into two separate likelihoods with different parameter sets. This likelihood factorization implies that maximum likelihood estimates of the main model are determined only from the corresponding likelihood part. For regression, it means only the maximum likelihood estimates of β-regression coefficients are needed. However, under MAR, the observed data are not a representative sample of the underlying population. This is important since in this case a simple mean of the observed data, $y_{i(obs)}$, will be a biased estimate (of the true parameter value). Unfortunately, it is not possible to test whether MAR is tenable compared to MNAR missingness mechanism explained next.

14.2.3 Missing Not At Random, nonignorable missing, or informative missing

Missing Not At Random (MNAR), also known as nonignorable missing or informative missing, is when the probability of missingness depends on both observed

and missing values. Under a MNAR process missingness relates not only to the previously observed disease state, but also to the unobserved missing disease state. This might occur for a condition with periods of quiescence and flare-ups such as temporomandibular joint and muscle disorder. Nonignorable models must assume a model for the missingness mechanism, i.e. for $P(r_i|y_{i(obs)}, y_{i(mis)})$. For a MNAR process, the observed responses are not representative of the full data and the joint distribution cannot be factorized. The implications to practice are very serious. Methods to correct for MNAR require specifying models for the probability distribution of missing value indicators in terms of the complete data, i.e. both observed and missing values. In other words, for MNAR no classical analyses or standard statistical software packages yield valid answers since the classical model needs to be combined with a model for missingness, and the two analyses cannot be fitted separately. On the other hand, when the correlation between unobserved data and observed data increases, e.g. in a longitudinal study when there is a high correlation between the subsequent values, then simulations have shown that there are fewer problems with MNAR (Molenberghs *et al*., 2008).

14.3 Approaches for analyzing missing data

14.3.1 Complete case (CC) analysis

The most common approach, *complete case* (CC) analysis, also known as casewise or listwise deletion, simply removes the participants who have any missing data. This is the default analysis in most statistical analysis software packages and the typical analysis reported in the oral health literature. There have been several rigorous investigations of the effects of CC analysis versus other approaches (Rubin, 1987; Molenberghs *et al*., 2004; Dmitrienko *et al*., 2005). The CC approach requires that missing values be MCAR and can be inefficient resulting in loss of power. Moreover, if data are not MCAR, substantial bias can result from CC analyses.

14.3.2 Imputation

14.3.2.1 Single, mean, hot decking, or Last Observation Carried Forward (LOCF) imputation

Imputation methods create or fill-in missing values and the literature contains many such approaches, although sometimes these may not be interpreted as imputation methods. Mean (or median) methods include imputing with the overall mean (median) value or the mean (median) value by cross-classified groups, say based on demographics. *Mean* value single *imputation* methods have the advantage of producing imputed data that preserve marginal means, i.e. the means for each variable where missing data are imputed remain the same. Drawbacks of this method are that associations (e.g. correlation coefficients) might be severely distorted. Also, it may result in imputed values that do not actually exist among any observed values or may not be plausible (e.g. fractional values for discrete integer counts). *Hot decking*,

in contrast, involves substituting one randomly selected actual observed value, either overall or by cross-classified groups. Hot decking which arose from the sample survey literature can yield means of the imputed data that differ from the CC data as expected under MAR missingness. Other single imputation methods, which may not have been considered imputation by their supporters, include *last observation carried forward* (LOCF), which had been advocated as a conservative ITT approach (Chapter 6) particularly for randomized trials. Since the most recent previous measure is often most correlated with the current value, LOCF fills-in missing data with the last observed measurement, which has been argued as being conservative since, for example in a therapeutic trial, illness from a prior period is assumed to still exist. While LOCF may initially seem like a reasonable method, recent rigorous evaluations (e.g. Molenberghs *et al*., 2004) have shown through derivations and simulations that the LOCF approach is biased in unknown ways, making stronger assumptions than even MCAR methods. Thus, LOCF should not be utilized.

All these approaches are single value imputation, which replace a missing datum with one value. Single value regression predictions typically underestimate variances (Allison, 2001), because they ignore the uncertainty in the imputed values. These problems with single value imputation have been studied extensively with one remedy being multiple imputation (Rubin, 1987; Schafer, 1997).

14.3.2.2 Multiple imputation

(MI) incorporates variability (uncertainty) of the imputation by repeating single value imputation M times with a random mechanism to create M different complete imputed datasets. The M datasets produce M point estimates and corresponding variance estimates, which are combined. The random mechanism is like taking random draws from the conditional distribution of the missing data given the observed data, which is why each of the M imputations differs slightly. Point estimates $\hat{Q}^{(1)}, \hat{Q}^{(2)}, \ldots, \hat{Q}^{(M)}$ and corresponding variance estimates $\hat{U}^{(1)}, \hat{U}^{(2)}, \ldots, \hat{U}^{(M)}$ are combined by averaging point estimates $\hat{Q} = \frac{1}{M}\sum_{m=1}^{M} \hat{Q}^{(m)}$ and calculating corrected precision $T = \hat{U} + (1 + M^{-1})B$ from within-imputation variance, $\hat{U} = \frac{1}{M}\sum_{m=1}^{M} \hat{U}^{(m)}$ and between-imputation, $B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}^{(m)} - \hat{Q})^2$ for a univariate estimate or $B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}^{(m)} - \hat{Q})(\hat{Q}^{(m)} - \hat{Q})'$ for a multivariate estimate (Rubin, 1987). The number of imputations, M, needed is surprisingly small. MI relative efficiency for M finite imputations relative to an infinite number of imputations is calculated as $1/(1 + \gamma/M)$, where γ is the fraction of missing information, so at least 96 % efficiency results from $M = 3$ with 10 % missing information, from $M = 10$ with 30 % missing information, and from $M = 20$ with 50 % missing information.

To be successful MI should employ all available knowledge of the subject with missing information. Thus, MI is based on rich statistical models involving all variables related to missingness and all covariates and responses for the analyses of interest; e.g. MI should include interactions and nonlinear terms that will be used in the analyses of interest. For sample surveys with complex design features, MI should also include survey design variables (Schafer, 1999; Schenker *et al*.,

2006; Reiter *et al*., 2006). Hence, MI assumes missing data are MAR; although MI cannot technically handle MNAR missingness mechanisms, every MNAR model can be fitted as a MAR model (Molenberghs *et al*., 2008). MI can handle arbitrary pattern MAR data (i.e. not only monotone patterns). It is also related to the Bayesian framework for Markov chain Monte Carlo (MCMC) models (see Chapter 18) where one alternates between an imputation step and a posterior probability draw (simulation) step until the models converge. To date, MI has been used on a limited basis in oral health research (e.g. Lento *et al*., 2004).

MI has been extended to multistage complex sample surveys with *sequential regression multivariate imputation* (SRMI) (Schenker *et al*., 2006) and associated free software (IVEware 2.0). In SRMI, cluster samples with both cluster-level and within-cluster-level covariates are first imputed at the cluster-level and then at the within-cluster-level for the complete survey. Then, like standard MI, this entire imputation sequence is repeated M times. SRMI uses MCMC, partitioning the joint conditional model and modeling each conditional density with a regression appropriate for the data type (e.g. linear, logistic, Poisson), drawing from each posterior predictive missing value distribution given the observed values.

14.3.3 Likelihood-based models

Under the MAR ignorability assumption, straightforward likelihood-based or direct likelihood analyses such as linear, generalized linear (including logistic regression), maximum likelihood, linear mixed, and generalized linear mixed models provide consistent estimates of quantities of interest. Direct maximization of the observed likelihood can be used. For example, with repeated continuous scale measures in two groups, the full general linear mixed effects model with the group x time interaction in the model provides proper inference (Molenberghs *et al*., 2004; Dmitrienko *et al*., 2005). Further details on these models can be found earlier in Sections 11.3–11.4, 11.6.2, and 13.2.3 (in contrast to GEE analysis in Chapter 13).

14.3.4 Weighted models (weighted generalized estimating equations)

An alternate strategy to accommodate MAR missing values involves inverse probability of selection methods from survey sampling. The marginal GEE models from Chapter 13 weight individuals by the inverse of the cluster size ($1/n_i$). Ordinary GEE methods can be extended to include weights based on the inverse probability of dropping out so that *weighted GEE* (WGEE) methods no longer require MCAR missingness for consistent estimation. WGEE methods only require MAR missing values (Robins *et al*., 1995). Weights for each subsequent response are conditional on not having dropped out previously. Weights are easily incorporated into standard GEE software. WGEE results typically increase standard errors, but since they are adjusted for missingness, they can account for bias and parameter estimates can change substantially. This is the tradeoff for GEE analyses that can accommodate data that are MAR.

14.3.5 MNAR (nonignorable) models

The aforementioned ignorable models need not model the missing data mechanism. In contrast, missing not at random (MNAR), or nonignorable, models must explicitly model the missingness mechanism, but these model assumptions cannot be assessed with observed data. MNAR models generally model the data and missingness mechanism jointly. Selection models and pattern mixture models are two general classes of models which can be used for MNAR situations (Dmitrienko *et al*., 2005). In *selection models* the joint distribution is factorized as $P(r_i, y_{i(obs)}, y_{i(mis)}) = P(r_i|y_{i(obs)}, y_{i(mis)})P(y_{i(obs)}, y_{i(mis)})$, while in *pattern mixture models* the joint distribution is partitioned as $P(r_i, y_{i(obs)}, y_{i(mis)}) = P(r_i)P(y_{i(obs)}, y_{i(mis)}|r_i)$. Pattern mixture models are sometimes more tractable than selection models. Pattern mixture models group subjects by their patterns of missing data (e.g. complete cases, missing the last measure, missing the last 2 measures, etc.). Some texts (e.g. Allison, 2001; Schafer, 1997) devote little attention to these models. Other authors (e.g. Dmitrienko *et al*., 2005) warn that care is needed in interpreting MNAR selection models especially with large sample sizes because of their strong untestable assumptions and recommend them for sensitivity analyses, i.e. a set of analyses varying model assumptions for the missingness mechanism.

14.3.6 Examples

14.3.6.1 Fluoride varnish early childhood caries prevention trial

In the prevention trial (Weintraub *et al*., 2006), the missing data pattern at 2-year follow-up related to observed caries status at 1-year follow-up ($p<0.001$) indicating data were not MCAR. Thus, MAR−compatible MCMC (for arbitrary/non-monotone missingness) MI with $M = 20$ imputations was used with the variables listed in Table 14.2. to impute log(1+number of decayed or filled tooth surfaces) scores, which were then dichotomized (dfs>0) for incidence analyses. This MI includes treatment group indicators as design variables (Allison, 2001; Schafer, 1999); others advocate separate MIs for each treatment group (Molenberghs *et al*., (2004, 2008)) but with these data the separate MI approach yielded similar results (not shown). MI diagnostic autocorrelation and time series plots of iterations were examined to assess model compatibility (Schafer, 1997). Various models (linear, log-linear, Poisson, overdispersed Poisson, zero-inflated Poisson, and negative binomial) were considered for number of decayed or filled tooth surfaces (dfs), with natural logarithm of 1+dfs having easiest interpretability among models with the best fit (smallest Akaike Information Criterion) (Gansky *et al*., 2005). Analyses were performed with SAS 9.1.2 PROC MI and PROC MIANALYZE. Bivariate binary GLMM for the 2 follow-up times with group, time and group x time effects was fitted with SAS PROC GLIMMIX for a direct likelihood approach. Results in Table 14.2 show that children missing follow-ups were projected by MI to have higher caries incidence, but for these data CC analyses agreed reasonably well with MAR-compatible GLMM analyses and MI

Table 14.2 Fluoride varnish randomized clinical trial complete case ($n = 280$) versus multiple imputation ($n = 376$) and direct likelihood (GLMM) analyses, logistic regression coefficients (standard errors) for treatment effects on caries incidence (positive number of decayed or filled tooth surfaces).

	CC	$MI-1^{\dagger}$	$MI-2^{\dagger}$	*GLMM*
dfs>0	28.2 %	38.3 %	37.4 %	–
Intended				
0v4	−1.33 (0.36)	−0.75 (0.31)	−0.85 (0.29)	−1.48 (0.39)
0v2	−0.79 (0.31)	−0.61 (0.29)	−0.54 (0.28)	−0.74 (0.30)
Actual				
0v3+	−2.91 (1.04)	−3.30 (1.04)	−3.20 (1.04)	−2.95 (1.03)
0v2	−1.23 (0.40)	−1.63 (0.39)	−1.53 (0.39)	−0.85 (0.34)
0v1	−0.91 (0.33)	−0.77 (0.29)	−0.62 (0.28)	−0.78 (0.31)

† variables for MI-1 included centre (strata), randomized treatment (intended) group indicators (2FV or 4FV), number of FV applications received (actual) indicators (1, 2, 3+), baseline factors related to loss-to-follow-up (mother's age, dental pain barrier, dental fear barrier, fluoride toothpaste use), and baseline log number of teeth (as a measure of dental age), while MI-2 also added baseline salivary bacteria measures ($\log_{10}$(mutans streptococci) and $\log_{10}$(lactobaccllli)).

analyses, with some attenuated effect on incidence. Similar findings resulted from linear models of log(1+dfs).

14.3.6.2 California Oral Health Needs Assessment of Children (2004-5) complex sample survey

In the COHNAC, school-level and examiner provided child-level covariates can help with multiple imputation since, for example, examiner rated race/ethnicity relates to parent reported race/ethnicity and school-level percent free/reduced-price lunch (FRL) program participation relates to child-level FRL participation. Although there is a high percentage (53 %) of missing covariate data, imputation methods may reduce bias compared to typical CC analyses. Any analyses, including missing data methods, should properly account for the complex survey design and survey sampling weights. As with SRMI for the National Health Interview Survey (Schenker *et al*., 2006), we implemented SRMI for COHNAC using IVEware, first to impute school-level information and then child-level variables with 10 imputation cycles, repeating the full imputation process independently 10 times (i.e. $M = 10$) for an estimated 94–95 % efficiency. IVEware used a multivariate Normal approximation of the posterior distribution and predicted values from the regression model conditional on perturbed coefficients. Variables included untreated caries, any caries experience, rampant caries, treatment urgency, lack of dental sealants (among third graders), last dental visit, age, gender, examiner-reported race/ethnicity, language used at home,

parent-reported race/ethnicity, insurance, inability to obtain dental care in the prior year, FRL program participation, barriers to obtaining dental care, school-level percent FRL program participation, average kindergarten-third grade class size, average fourth-sixth grade class size, school-level parental education distribution, school year type (year round or traditional), school California Dental Disease Prevention Program participation, school consent type (active or passive), dentist to population ratio in the school's rational care area, additional census and school information, and survey design variables such as indicator variables for strata and clusters (Reiter *et al*., 2006). Survey weights were incorporated and a restriction statement was used for skip patterns (e.g. sealants only apply to third graders) as well as a bounds statement to limit imputation to legitimate ranges.

Estimated proportions and standard errors from unimputed (CC) and MI data are shown in Table 14.3. Modest differences occurred in parent report of Hispanic ethnicity, insurance status and participation in the school's FRL program, although standard errors are similar. Importantly, the impact is not only in the sample characteristics, but also in model results as shown in the survey logistic model of untreated caries using CC, naïve single imputation and MI in Table 14.4. Here, both imputations produced large precision gains with much smaller standard errors, as expected by the increased sample size. Moreover, CC model coefficients compared to MI coefficients were underestimated by as much as 40 % (English as a second language) and overestimated by as much as 44 % (no insurance), illustrating that the CC analyses can be substantially biased in either direction. Although naïve single imputation may appear better than CC in this particular situation, it ignores uncertainty in imputation, underestimates SEs, and is not as reliable as MI; single imputation is not a recommended viable approach.

Table 14.3 Characteristics of COHNAC, unimputed and multiply imputed (n = 21,399).

Variable	*% Missing*	*Unimputed (CC)*		*Multiply Imputed*	
		Proportion	*SE*	*Proportion*	*SE*
Hispanic ethnicity	54	0.521	0.031	0.558	0.029
African-American race	54	0.058	0.007	0.052	0.006
Asian race	54	0.116	0.013	0.109	0.011
Free lunch program	57	0.473	0.025	0.490	0.023
Private insurance	67	0.375	0.024	0.385	0.020
Government insurance	67	0.389	0.020	0.407	0.019
No insurance	67	0.235	0.008	0.207	0.005
Male gender	0.6	0.498	0.005	0.499	0.005
English 2nd language	0.8	0.415	0.033	0.415	0.032
Any caries experience	0.1	0.626	0.011	0.626	0.011
Untreated caries	<0.1	0.283	0.009	0.283	0.009
Rampant caries	0.1	0.207	0.011	0.208	0.011
Urgent treatment need	0.2	0.043	0.005	0.043	0.005

Table 14.4 Logistic regression model of untreated caries in COHNAC.

Covariate	*Unimputed (CC) (n=6149)*		*Singly Imputed (n=21,399)*		*Multiply Imputed (n=21,399)*	
	Beta	*SE*	*Beta*	*SE*	*Beta*	*SE*
Hispanic ethnicity	0.240	0.113	0.246	0.058	0.234	0.059
English 2nd language	0.162	0.104	0.232	0.062	0.227	0.060
Free lunch program	0.342	0.091	0.270	0.050	0.273	0.050
Government insurance	0.329	0.096	0.223	0.048	0.268	0.059
No insurance	0.722	0.094	0.366	0.054	0.405	0.067

14.4 Informative/nonignorable cluster size

Other issues specific (although not unique) to oral health research can impact missing data. In particular, issues related to an individual's number of teeth (cluster size) can affect statistical analysis. Disease itself and other factors such as age and socioeconomic status often relate to an individual's number of teeth. There are many circumstances in nature in which the missing within-cluster elements (i.e. cluster size) relate to the underlying health status: gravidity (number or pregnancies) inversely with parity (number of births) among fertility clinic couples trying to conceive who continue conception attempts and pregnancies after unsuccessful deliveries (Hoffman *et al*., 2001); family (litter) size negatively with toxicity (Hoffman *et al*., 2001; Dunson *et al*., 2003); cattle herd size positively with infection rate since infection can spread faster in higher densities (Boelaert *et al*., 2005); clinic size negatively with phototherapy for newborns with high bilirubin since facilities with many patients might not have enough phototherapy units (Neuhaus & McCulloch, 2006); and condom use negatively with frequency of encounters (Williamson *et al*., 2007). In oral health studies, number of teeth present may be negatively related to periodontal health (Hoffman *et al*., 2001; Williamson *et al*., 2007 and cleft palate incidence may negatively relate to dexamethasone exposure (Chen, 1993).

This phenomenon is shown in Figure 14.2 with data from the intergenerational epidemiologic cohort study of adult periodontitis (Gansky *et al*., 1999), plotting number of posterior teeth (cluster size) versus the proportion of posterior teeth with worst loss of attachment (WLA) per tooth ≥ 3 mm using nonparametric loess smoothing to illustrate the relationship. Additional examples could include: the number of pain episodes positively related with diary pain scores or the number of visits attended in a longitudinal clinical trial positively related to health, due to a type of 'healthy worker effect' since trial participants are usually healthier than non-participants.

Earlier simulation studies (Williamson *et al*., 2003; Neuhaus & McCulloch, 2006) have concluded considerable bias in mean, proportion, and intercept estimates of regression models can result from informative cluster sizes. However, recent studies show little bias in slope (association) regression estimates under informative cluster size. For example, linear mixed effect model slope coefficients are unbiased

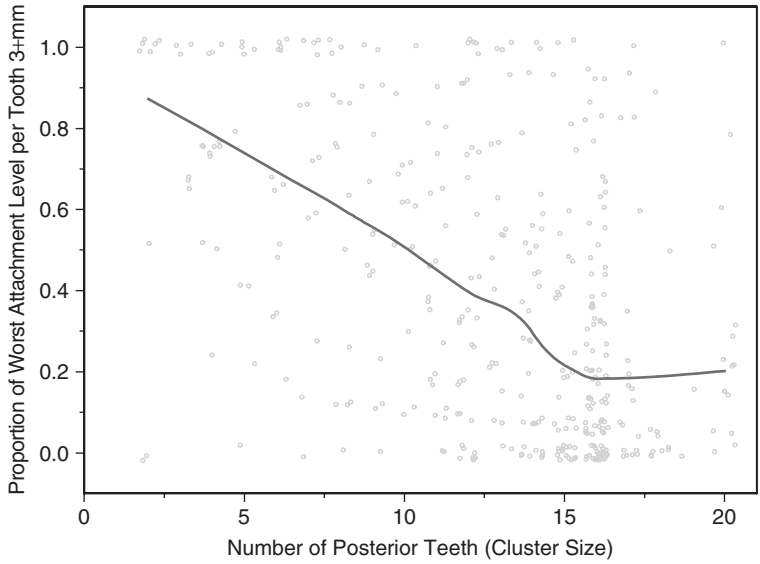


Figure 14.2 Informative cluster size: Cluster size (n_i) relates to adult periodontitis status (n=406) (jittering with LOESS smoothing).

(Benhin *et al*., 2005). Moreover in GLMMs, ignoring informative cluster size can be viewed as a misspecified random effects distribution, having little effect upon slope estimation (Neuhaus & McCulloch, 2006). Bias can be somewhat alleviated with simpler models (partitioning or conditional ML) than previously advocated (i.e. resampling and weighted GEE (WGEE) (Robins *et al*., 1995).

In the presence of *informative or nonignorable cluster size*, the typical conditional expectation no longer holds true: $E(y_{ij}|x_{ij}, n_i) \neq E(y_{ij}|x_{ij})$, where y_{ij} is the outcome variable and x_{ij} is the explanatory variable for the ith subject's jth tooth, and n_i is the within-subject cluster size (e.g. number of teeth). Several approaches have been utilized including marginal methods of computationally intensive resampling (Hoffman *et al*., 2001) and WGEE (Williamson *et al*., 2003; Robins *et al*., 1995). However, other models are appropriate, as well. *Conditional maximum likelihood* (CML) canonical link models applied to overcome possible problems from informative cluster size (a special case of random effects being correlated with covariate distributions) condition on the cluster size and the sufficient statistic of the number of within-cluster events to remove the random effects and produce consistent slope estimates; generalized linear mixed models (GLMMs) partitioning covariate effects into *within- and between-cluster covariate components* have been shown theoretically and asymptotically to approximate CML models (Neuhaus & McCulloch, 2006). First the covariate effect, $x_{ij}\beta$, is decomposed into $\bar{x}_i\beta_B + (x_{ij} - \bar{x}_i)\beta_w$, where $\bar{x}_i$ is the within-cluster mean so $x_{ij} - \bar{x}_i$ is the within-cluster deviation from the mean; then the GLMM is fitted. A recent paper suggests using a mixed effects logistic GLMM with cluster size as a covariate

Table 14.5 Informative cluster size model effect differences for within-cluster covariate (molar).

Parameter estimated	*Method*				
	logit GLMM	*logit GLMM* $y_{ij}\|x_{ij}, n_i$	*B/W logit GLMM*	*CML*	*Joint* $y_{ij}, n_i\|x_{ij}$
β_0	−1.328 (0.153)	3.245 (0.416)	1.951 (0.443)	–	−1.208 (0.149)
β_{Molar}	0.787 (0.090)	0.827 (0.090)	–	0.843 (0.090)	0.842 (0.090)
β_B	–	–	0.841 (0.090)	–	–
β_W	–	–	−6.781 (0.983)	–	–
n_i	–	−0.379 (0.034)	–	–	–
$\log(\sigma_b^2)$	0.982 (0.056)	0.784 (0.057)	0.890 (0.057)	–	0.960 (0.055)

to correct for informative cluster size (Chen, 1993; Faes *et al*., 2006), although unpublished simulations (Neuhaus) have shown that method to still be biased. (See also Chapter 13.)

We examined the performance of several approaches to accommodate informative cluster sizes using data from the intergenerational epidemiologic study of adult periodontitis. The analyses examined the relationship between posterior tooth type (molar versus premolar) and worst loss of attachment per tooth ≥ 3 mm (Gansky *et al*., 1999) fitting a standard mixed effects logistic GLMM, a mixed effects logistic GLMM with cluster size as a covariate, a mixed effects between- and within-cluster partitioned logistic GLMM, and a joint model of WLA ≥ 3 and cluster size with SAS 9.1.2 PROC NLMIXED, while a CML model was fitted with SAS 9.1.2 PROC PHREG (as in Neuhaus & McCulloch, 2006). Results as shown in Table 14.5 illustrate the conclusions of the recent theoretical and simulation results (Neuhaus & McCulloch, 2006): slope, β_{Molar}, from the standard logistic GLMM was underestimated (compared to the joint model) by 6.5 % and from the logistic GLMM with number of posterior teeth as a covariate by <1.8 %, while between- and within-cluster partitioned logistic GLMM and CML models differed by <0.2 %, basically unbiased. Estimates of intercept, β_0, were biased for the logistic GLMM adjusting for number of posterior teeth as a covariate and for the between- and within- (B/W) cluster partitioned logistic GLMM. Thus, for correct inference of within-cluster covariate parameters estimated (e.g. β_{Molar} and β_B), the between-within cluster partitioned logit GLMM, CML, and joint model are all viable choices.

14.5 Summary and recommendations

We have introduced and described various common scenarios of missing data in oral health studies as well as typical and more recent analytic approaches. We recommend studies plan and organize to reduce missing data (see Chapters 4–7 and 9). When faced with missing data, ad hoc analytic approaches (including complete cases (casewise deletion), last observation carried forward, and single imputation) should be avoided. Instead, direct likelihood methods (e.g. maximum likelihood

estimation and generalized linear mixed models) and multiple imputation methods are recommended since they are appropriate not just for data missing completely at random but also for data missing at random (MAR). Weighted GEE models also are appropriate for MAR data, but require fitting the correct weighting model. If missing not at random models are used, sensitivity analyses are recommended since the assumed missing data model cannot be practically assessed.

14.5.1 Acknowledgments

The FV RCT was supported by US DHHS NIH/NIDCR P60 DE013 058 and NIDCR&NCMHD U54 DE014521. The COHNAC 2004-5 was performed under the auspices of the Dental Health Foundation with support from the US Health Resources and Service Administration, the California Dental Association Foundation, the First 5 California, the California Endowment, and the Association of State and Territorial Dental Directors. The Intergenerational Epidemiologic Study of Adult Periodontitis was supported by US DHHS/NIH/NIDCR R01DE09856. We thank Drs Bahjat Qaqish and John Preisser, University of North Carolina at Chapel Hill, for reporting apparent identification number mislabeling in a prior version of the periodontitis dataset, which is now corrected. This work was supported in part by US DHHS/NIH/NIDCR R03 DE018116. We thank Dr Sara Shain and Ms Nancy Fan Cheng, University of California, San Francisco and the editors and statistical and dental research reviewers for their comments on an earlier version of this Chapter; any errors or omissions that remain are solely our own.

References

Allison, P. (2001) *Missing Data*, Sage Publications, Thousand Oaks, CA.

Benhin, E., Rao, J.N.K., & Scott, A.J. (2005) Mean estimating equation approach to analyzing cluster-correlated data with nonignorable cluster sizes. *Biometrika* **92**: 435–50.

Boelaert, F., Speybroeck, N., & Kruif, A. d. *et al*. (2005) Risk factors for bovine herpesvirus-1 seropositivity. *Preventive Veterinary Medicine* **69**: 285–95.

Chen, J. (1993) A malformation incidence dose-response model incorporating fetal weight and/or litter size as covariates. *Risk Analysis* **13**: 559–64.

DeLeeuw, E.D. (2001) Reducing missing data in surveys: An overview of methods. *Quality and Quantity* **35**: 147–60.

Dental Health Foundation (2006) California Smile Survey: An oral health assessment of California's kindergarten and 3rd grade children – Methods, Oakland, CA, http://www.dentalhealthfoundation.org/topics/public/For%20web/I.CA%20Smile%20Survey%20Methods.pdf.

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C. & Offen, W. (2005) Analysis of incomplete data. In: *Analysis of Clinical Trials Using SAS: A Practical Guide*, SAS Institute Inc, Cary, NC, pp. 259–354.

Dunson, D.B., Chen, Z., & Harry, J. (2003) A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59**: 521–30.

Faes, C., Hens, N., Aerts, M. *et al*. (2006) Estimating herd-specific force of infection by using random effects models for clustered binary data and monotone fractional polynomials. *Applied Statistics* **55**: 595–613.

Gansky, S.A., Weintraub, J.A., Shain, S. & Multi-Pied-Investigators (1999) Family aggregation of periodontal status in a two-generation cohort. *Journal of Dental Research* 78(Spec Iss B): 123.

Gansky, S.A., Shain, S.G. Ramos-Gomez, F., & Weintraub, J.A. (2005) Fitting statistical models for early childhood caries. *Journal of Public Health Dentistry* S64.

Hoffman, E.B., Sen, P.K., & Weinberg, C.R. (2001) Within-cluster resampling. *Biometrika* **88**: 1121–34.

Lento, J., Glynn, S., Shetty, V., Asarnow, J., Wanq, J., & Belin, T.R. (2004) Psychological functioning and needs of indigent patients with facial injury: A prospective controlled study. *Journal of Oral Maxillofacial Surgery* **62**: 925–32.

Little, R.J.A. & Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York.

Molenberghs, G., Thijs, H., Jansen, I. *et al*. (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5**: 445–64.

Molenberghs, G., Beunckens, C., Sotto, C. & Kenward, M.G. (2008) Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B* **70**: 371–88.

Neuhaus, J.M. & McCulloch, C.E. (2006) Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society B* **68**: 859–72.

Reiter, J.P., Raghunathan, T.E., & Kinney, S.K. (2006) Importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology* **32**: 143–9.

Robins, J.M., Rotnitzky, A., & Zhao, L.P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **93**: 1321–39.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc., New York.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, New York.

Schafer, J.L. (1999) Multiple imputation: A primer. *Statistical Methods in Medical Research* **8**: 3–15.

Schenker, N., Raghunathan, T.E., Chiu, P.L., Makuc, D.M., Zhang, G., & Cohen, A.J. (2006) Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* **101**: 924–33.

Weintraub, J.A., Ramos-Gomez, F., Jue, B. *et al*. (2006) Fluoride varnish efficacy in preventing early childhood caries. *Journal of Dental Research* **85**: 172–6.

Williamson, J.M., Datta, S., & Satten, G.A. (2003) Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59**: 36–42.

Williamson, J.M., Kim, H.Y., & Warner, L. (2007) Weighting condom use data to account for nonignorable cluster size. *Annals of Epidemiology* **17**: 603–7.

15

Failure time analysis

Thomas A. Gerds, Vibeke Qvist, Jörg R. Strub, Christian B. Pipper, Thomas H. Scheike and Niels Keiding

15.1 Introduction

In survival analysis one is interested in the probability that an event occurs before or after certain time points. For example, suppose that for several dental implant systems the probabilities are given that no adverse event occurs during the first five years. This information, possibly complemented by aesthetic or health-conscious aspects, supports the treatment decision and can easily be understood by the patient. Thus often the aim of the statistical analysis is to predict the survival chances of a tooth, a filling, an implant or a similar study unit. Other important parameters are regression coefficients that describe the influence of patient and tooth specific factors on the event probabilities, and further measures for the association between event times in the same mouth.

Two features complicating the analysis are common for applications of survival analysis to dental research: First, within the framework of a dental study it occurs naturally that the exact event times of some or all study units remain unknown to the data analyst. Besides patient withdrawal, the main reason is that typically the status of the study unit can only be evaluated when the patient is examined by the dentist. Secondly, and this is the most crucial difference to other fields of application, there is often an inherent **cluster-correlated** structure in the data: two study units placed in the same patient will rarely behave independently.

This chapter introduces the statistical concepts and illustrates adaptation of classical survival techniques to applications in dental research. However, due to

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

a lack of methodology and software it will often not be possible to handle all complications at the same time; for example when a regression analysis has to be based on interval censored cluster-correlated event times in the presence of competing risks. We get back to the feasibility issue in the last section.

15.2 Statistical concepts

To study the performance of dental treatment it is often useful to define events that characterize the success or the failure of the study units. Survival analysis or failure time analysis, also called event history analysis, describes the probabilities with which the events occur to a study unit in time after treatment. To estimate the probabilities all relevant events are recorded and, when it is possible, also their onset times.

In classical survival analysis there is only one relevant event: the failure of the study unit. The main tool to describe event probabilities in classical survival analysis are the **survival function** and the **hazard rate**. The value of the survival function at a given time point is interpreted as the probability that a study unit survives this time point. The value of the hazard rate at a given time point is interpreted as the probability that the study unit fails precisely at that time point given that it has survived up to that time point. If the study unit has not survived then the hazard rate is zero. The aim is to estimate the survival function and the hazard rate from data. To a great extent the estimation is involved because typically it is not possible to observe the exact failure times of all study units.

15.2.1 Right censored data – Kaplan-Meier method

Event times are called **right censored** if the event did not occur during the study period. For those study units the end of the study period is called the **censoring time**. Right censoring occurs for example when the patient is lost to follow-up, or withdrawn from the study for some reason, or simply when the end of the study time is reached for the study unit and the unit is still in function. For right censored event times the Kaplan-Meier method can be used to estimate the survival function.

In a clinical cohort study [1] 211 dental implants were placed in 61 patients and then followed for several years with recording of failures. Between 2 and 10 study implants were placed in each patient. A typical survival analysis could address the 5-year survival probability for the TPS-SteriOss® implants that were used in this study. In the current example the event is failure of implant. Overall 13 implants failed during their study period. For the remaining 198 implants the failure times are right censored. The idea of the Kaplan-Meier method (see e.g. [2] for a formal description) is to not ignore the right censored observations, but to consider the possibility that some of these (right censored) implants could have failed before 5 years (only this has not been observed in the study).

The Kaplan-Meier curve is a step function with jumps at all failure times. In case of uncensored data the jump size equals $1/n$ where n is the sample size;

then the Kaplan-Meier method yields exactly the same as the empirical distribution function. If, however, some **failure times** are right censored then the Kaplan-Meier method systematically increases the jump size to account for that some of the earlier right censored implants might have failed. Applied to the data of the TPS-SteriOss® implant study the Kaplan-Meier method yields a 5-year survival probability of 93.4 %.

It is important to recognize that simple descriptive statistics are not to be used in this context because of right censoring. To make this point clear consider the crude failure rate at 5 years. In this study 11 of the 13 failures occurred before 5 years. If instead of using the Kaplan-Meier method one would use the crude rate, that is 11 divided by 211, to estimate the failure rate, one would get to a survival probability of 94.7 %. The crude rate ignores the right censored observations and it thus yields a too high survival probability. The crude rate produces systematically misleading results and should not be used with censored data.

In principle the survival function can also be estimated based on the right censored data by using a parametric model, for example a Weibull model [3]. However, a parametric model typically makes strong assumptions about the form of the survival function and the results can only be trusted if these assumptions are satisfied. The Kaplan-Meier method does not make parametric assumptions about the unknown survival probability distribution, it is a nonparametric method, and therefore often preferable. However, the Kaplan-Meier method relies on the following structural assumption regarding the censoring mechanism:

1. For each right censored study unit the event will occur later, only the event time remains unknown.

2. The survival probabilities do not change due to censoring. In mathematical terms: the censoring time is independent of the event time. In practical terms: The likelihood that an implant fails for a patient who has been lost to follow-up is the same as the likelihood for a patient who is still under study.

The first assumption will be discussed in detail in Section 15.2.4. If the second assumption is satisfied, then the performance of the right censored study units at times after their censoring time is comparable to the performance of the study units still under observation. Thus it is possible to borrow the information from the study units that are not censored in that period without introducing a bias. More details and corresponding assumptions for more complex censoring patterns can be found in [4] and [5].

15.2.2 Left censoring

Left censoring occurs when events have happened before some time, but we do not know when. A possible example in dental research could be a follow-up study determining the age of exfoliation of primary teeth. If this study starts at age 8

there may be teeth that were already exfoliated before the study started, and we do not know when.

15.2.3 Interval censored data

In a prospective cohort study [6] the investigators followed 529 amalgam and 471 glass ionomer cement fillings in primary molars. The study recruited children at their regular visit to one of 14 dentists in several Danish municipalities. In this example the study unit is the filling. Between 1 and 7 fillings were placed in each child. The performance of the fillings was assessed at subsequent routine visits. During the study period 146 fillings fractured and it is of interest to compare the fracture probabilities of the two materials. It has to be noted that in this example the fracture probabilities cannot be estimated with the Kaplan-Meier method. One reason is that the exact onset times of filling fracture are unknown because fracture of filling has to be diagnosed by the dentist (a second reason is explained in Section 15.2.4).

In general, event times are called interval censored [7] when it is only known that the event occurred between two adjacent examination times. This occurs for example when an event can only be detected by the dentist and thus the state of the study unit is unknown in the period between the examination times.

In the current example 146 fillings fractured during their study period and the fracture times are interval censored (Figure 15.1). Since in this study the examinations were not scheduled, and not all children went regularly to see the dentist, the observed intervals of fracture times have rather unequal length (median 191 days; min 0 days (exact time); max 813 days).

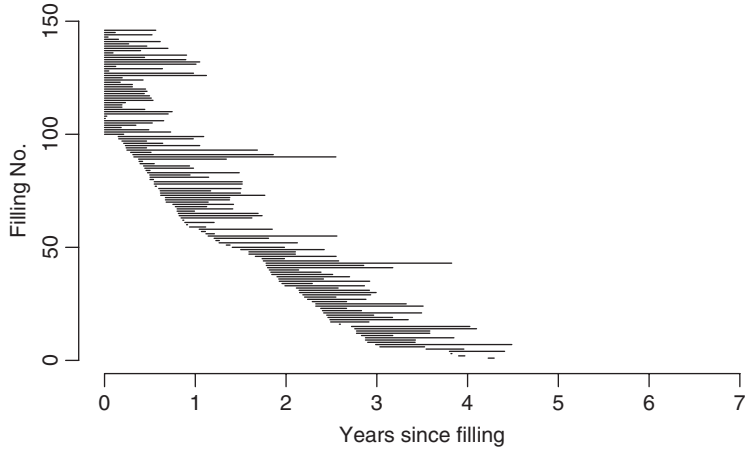


Figure 15.1 Interval censored observation of 146 fracture times as observed for fillings in the primary teeth study. Each black line represents an interval bounded by the date of the first examination where the filling was diagnosed as fractured and the date of the preceding examination.

The fracture probabilities of the two materials can be estimated based on the interval censored data with advanced statistical methods (see Section 15.3.2).

A special form of interval censored data occurs when only one examination is performed per study unit. If the event can be diagnosed at this examination it is actually only known that the event time is smaller than the examination time. Otherwise it is only known that the event time is larger than the examination time. For example, in a cross-sectional study of first caries in the primary dentition the status is recorded together with the age of each individual at a single examination. Thus the age-specific presence of caries in a cross-sectional sample may be converted to an age-distribution of first caries. The data arising from such a study are called type-1 interval censored or current status data [8].

15.2.4 Competing risks

A **competing risk** is an event after which the study unit is not exposed to other events. Competing risks are also called **causes of failure**. In a situation with competing risks event probabilities can be estimated with the Aalen-Johansen method. For details see [2, Chapter 10] and also Section 15.3.2 below.

Tooth exfoliation is the physiologic loss of the primary dentition. In the study [6] most of the primary teeth exfoliated during the observation period before the filling failed. One may be tempted to treat tooth exfoliation as study end without failure, that means as right censored, and e.g. use the Kaplan-Meier method to estimate the failure probability of the fillings. However, a filling cannot fracture after exfoliation of the associated tooth and thus exfoliation is a competing risk which takes the filling from being exposed. The fundamental difference between right censored and competing risks is explained for the fillings as follows: If the study period ends and the filling is still intact then the event time is right censored. Thus at the last examination (the time of right censoring) the filling is still exposed and may fracture later. However, a filling in a primary tooth is not exposed after the tooth has been replaced by a permanent tooth and thus exfoliation is not the same as right censoring.

Treating tooth exfoliations as right censored can lead to severe overestimates of the failure probability. This bias is illustrated in Figure 15.6 where the filling fracture probability is estimated. The dashed line is obtained with the Kaplan-Meier method and by treating tooth exfoliations as right censored. The dashed line lies far above the solid line which is obtained with the Aalen-Johansen method and by treating the tooth exfoliations as competing risks. See Section 15.3.2 for more details.

Another example in dentistry where competing risks are relevant is when studying the outcome of implant-supported fixed prostheses. Here for example death of patient and implant failure are relevant competing risks [9].

15.2.5 Marginal versus conditional modelling

The data arising from a typical dental study are cluster-correlated. Study units belong together in small groups. For example, when studying multiple fillings in

the same mouth, or multiple patients treated by the same dentist, or multiple events that occurred to a single study unit.

When considering several units there are two basic ways to model: the **marginal approach** and the **conditional approach**.

In the marginal approach the survival distribution of each unit is first modelled separately, and subsequently the possible dependence between units is handled. The marginal approach corresponds to focusing on the population of fillings, e.g. when the health system wants to compare longevity of dental filling materials in the population.

The conditional approach recognises the heterogeneity of patients first, usually through a distribution of a latent so-called **frailty** variable that affects all survival times in a patient similarly. The modelling is then specified in the conditional distribution given the frailty, and all model parameter estimates are to be interpreted for a given patient. This viewpoint corresponds to the patient asking about the relative benefits of various dental filling materials 'for me'.

15.2.6 Event history and hazard regression

The survival chances of a dental implant depend among other things on the implant length, the jaw site, on the smoking habits of the patient [10]. In general, the failure probability depends on what happened to the study unit so far and on factors that characterize the study unit and the patient. This information is called the **event history**. The hazard rate (see Section 15.2) always depends on the event history. In many applications a primary goal is to analyse how much the hazard rate depends on a specific factor. For this purpose one can use a **regression model**. In survival analysis regression models are usually formulated directly in terms of the hazard rate which may fully or partly depend on the information provided by the event history. The most prominent regression model for the hazard rate is the Cox regression model (Section 15.4).

15.2.7 Confidence intervals – bootstrap method

The results of a statistical analysis are estimates and therefore uncertain. For example, it is important to display the uncertainty of the Kaplan-Meier estimate for the 5 year survival probability of a dental implant system. Here the term uncertainty refers to the confidence with which any estimate can be accepted as representing the true performance of a dental treatment about which the analyst has incomplete knowledge. It is important to note that in this discussion uncertainty is not used as a synonym for risk. The term risk is interpreted to mean the probability of a particular event occurring. The uncertainty is introduced by sampling variation which will typically be lower when regarding the event times within the same patient (cluster) compared to when regarding the event times from different patients (clusters).

In a conventional survival analysis, it is common to derive confidence intervals for the Kaplan-Meier method from the Greenwood estimate of variance [5]. Consider the implant study described in Section 15.2.1. Of interest is the standard error

for the Kaplan-Meier estimate of the 5 year implant survival probability. Greenwood's formula yields $se = 0.0203$. This estimate, however, does not account for the fact that multiple implants are placed in the same mouth and is typically too small. A more appropriate formula for the Kaplan-Meier method with cluster-correlated event times is described in [11, 12] and its relevance for dental implant studies is illustrated in [13]. In the current example the modified formula yields $se = 0.0292$, and derived from this, a 95 % confidence interval for the 5 year survival probability of TPS-SteriOss® implants is [89.4 %–97.3 %].

In general, when the event times are cluster-correlated the usual variance estimators have to be modified. Reference [14] describes a relatively simple recipe for this. Alternatively one may obtain standard errors via a computer based resampling procedure called the bootstrap [15]. The advantage of this procedure is that it does not require a formula for the standard error. The bootstrap for cluster-correlated data works as follows. Step 1: Draw (on the computer) clusters (i.e. patients) with replacement from the original sample. Step 2: Compute the estimate, for example the 5 year survival probability, in that newly created sample. Step 3: Repeat steps 1-2 many times (500 to 10000) to get a large vector of estimates. Step 4: Use the sample variation of the data obtained in Step 3 to approximate the standard error.

15.3 Estimating event probabilities

15.3.1 The survival function

The classical survival model uses only two different states to describe the course of dental treatment. It is most common that all study units start in the first of the two states, usually an initial state of comfort and function, and stay there until the event of interest occurs. Often this event is associated with failure of the study unit as in Figure 15.2. We use the index *'ik'* for the *'k'*th study unit belonging to patient *'i'*. The two-state survival model can be parametrised alternatively through the **failure probabilities**

$$F_{ik}(t) = \text{Prob}\ (\textit{Study unit 'ik' is in the failure state at time } t)$$

the survival probabilities

$$S_{ik}(t) = \text{Prob}\ (\textit{Study unit 'ik' is in the initial state at time } t) = 1 - F_{ik}(t)$$

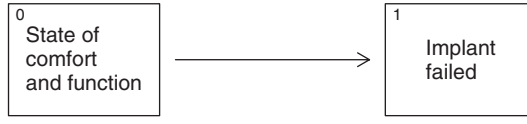


Figure 15.2 Two-state survival model for implant failure. The failure probabilities are described by the transition intensity between state 0 and 1.

or the hazard rate

$$\lambda_{ik}(t) = \text{Prob}\ (\textit{Study unit 'ik' fails at t, given it survived until t}) = \frac{f_{ik}(t)}{S_{ik}(t)},$$

where $f_{ik}(t)$ is the probability density corresponding to F_{ik}.

As noted in the introduction of this chapter it may be of interest to know the survival probability for a given implant system. Taking the marginal approach described in Section 15.2.5, the Kaplan-Meier method can be used to estimate a survival probability which is meaningful for the population of implants.

Figure 15.3 shows the estimated survival probability for TPS-SteriOss® implants at all time points between implantation and 6 years. In principle the survival function can be estimated until the last implant is either right censored or has failed. Note that the figure as given is not standard; it differs from the usual Kaplan-Meier plot in the following ways:

1. The current numbers of patients (clusters) are given for whom at least one implant (study unit) is uncensored, i.e. still under observation, and has not yet failed. This information complements the usually given number of uncensored implants (study units) that have not yet failed. The information would certainly be quite different if for example the 76 implants observed at risk of failure at 5 years would belong to 50 instead of 22 different patients.

2. The error bars represent pointwise 95 % confidence intervals obtained with the formula for cluster-correlated data (Section 15.2.7). It might be

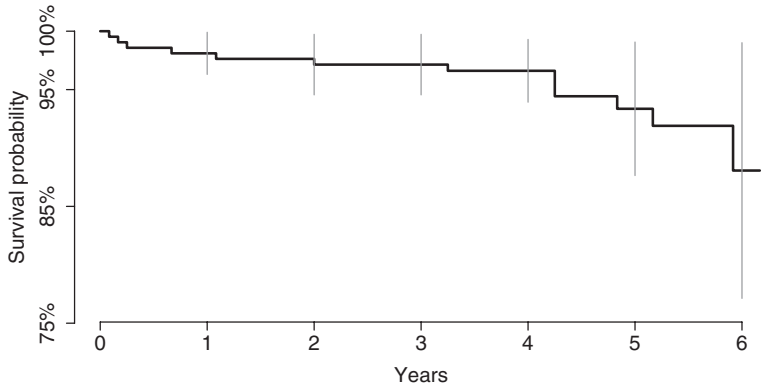


Figure 15.3 Kaplan-Meier estimate of the survival probability of the dental implants. Below the figure: the number of study implants that have not failed yet and are still under observation and the number of patients contributing with at least one study implant. The vertical gray bars represent pointwise 95% confidence limits.

surprising that after 6 years the estimated survival chance is about 80 % although only 16 of the 211 implants are still under observation and have not failed. Note that the low number of implants observed at risk in late parts of the time axis, e.g. only 16 implants in 4 patients after 6 years, is mostly explained by loss to follow-up and not by failures.

15.3.2 The cause-specific cumulative incidence

For some applications the two state model is too simple. An important extension of the two-state survival model is the competing risk model. It describes situations where some study units never fail from the relevant cause (see Section 15.2.4).

For example to assess the performance of fillings in primary teeth it is of interest to estimate the probability that a filling fractures. This probability is increasing with time and called the cumulative incidence of fracture. When there are competing risks, this probability cannot be obtained by subtracting the survival probability from 1.

In general, the **cumulative incidence function** describes the probability that a study unit fails from a specific cause of interest until time t.

In the primary teeth study 146 fillings fractured, 72 fillings were replaced because of endodontic complications and 96 due to other complications with the filling. In 561 cases the associated tooth exfoliated before any filling failure could be detected and for 125 fillings no event occurred until the end of the study time (Figure 15.4).

Unfortunately, in the current example the cumulative incidence of filling fractures has to be estimated from the interval censored data. But for the purpose of illustration it is possible to treat the midpoints of the interval censored times

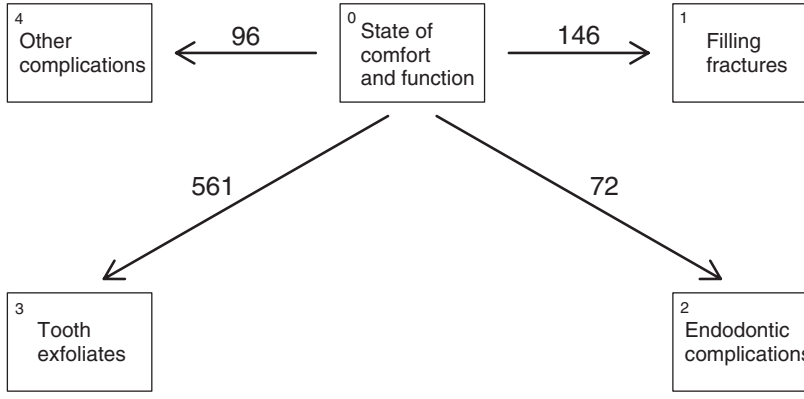


Figure 15.4 Competing risks model for fillings in primary teeth. After exfoliation of the tooth the filling is not at risk of failure anymore. Above each arrow is the respective number of fillings that left the initial state in this direction. The remaining 125 fillings were right censored before any of the displayed events occurred.

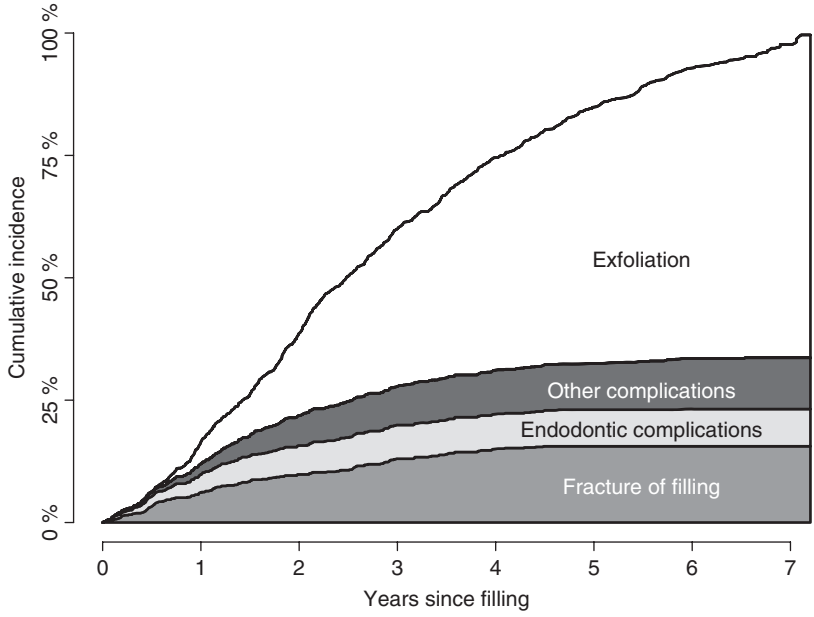


Figure 15.5 Cumulative incidence curves for the 1000 fillings in primary teeth corresponding to the model depicted in Figure 15.4: Fracture of filling ($n = 146$), endodontic complications ($n = 72$), other complications ($n = 96$), exfoliation of tooth ($n = 561$). The estimates are obtained with the Aalen-Johansen estimator and the midpoints of the interval-censored data.

as if they were the true event times, and then apply the Aalen-Johansen estimate which accounts for competing risks (Section 15.2.4). For example, by using this ad-hoc approach it is estimated that the probability that a glass ionomer cement filling fractures within the first three years is 23.9 % (95 %-confidence interval: 20.0 %–27.8 %). To show the effect of treating the event exfoliation and the other competing risks as end of study time the Kaplan-Meier method is also applied based on the midpoints of the intervals. This yields a probability of 32.6 % (95 %-confidence interval: 27.1 %–38.1 %) that a glass ionomer cement filling fractures within the first three years. Thus, in a hypothetical world where exfoliation of primary teeth does not occur the probability of filling fracture is considerably higher. In our context, however, this hypothetical world would rarely be relevant, because hypothetical 'fractures after exfoliation' would then be included in the analysis but without practical interpretation. The complete picture is shown in Figure 15.5.

The midpoint approach has been criticised because of its bias [16], which is specifically expressed when the observed intervals are wide, and also because the standard errors are too small [7]. A consistent estimate can be obtained with the so-called nonparametric maximum likelihood estimator for interval censored data [17; 18]. Like the Kaplan-Meier estimator the estimator results in a step function

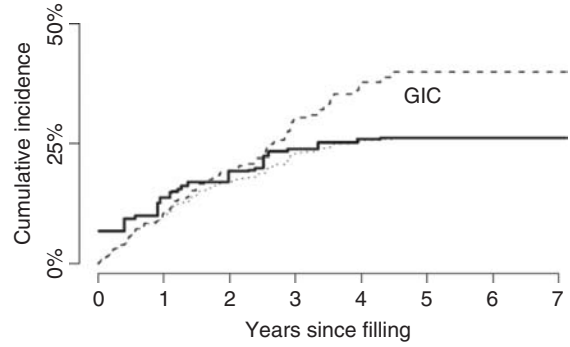


Figure 15.6 The cumulative incidence of filling fractures for the glass ionomer cement fillings. The solid line is obtained with the nonparametric maximum likelihood estimator for the competing risks model depicted in Figure 15.4. The dotted line is obtained with Aalen-Johansen estimator for the same model based on the interval midpoints and the dashed line is the Kaplan-Meier estimator based on the interval midpoints and treating competing risks as right censored.

that jumps only at selected times as described in [19]. Unfortunately there is no explicit formula to compute the estimate and it has to be obtained recursively with an algorithm [20; 21].

Most of the literature on interval censored data deals with the simple survival model. However, an extension of this method to estimate a cumulative incidence function from interval censored data has recently been developed [22]. For the purpose of illustration it is sufficient to consider the glass ionomer cement fillings. However, it shall be noted that the amalgam fillings performed much better with respect to fracture resistance [6]. Figure 15.6 illustrates the two biases that are discussed in this section.

The solid line represents the nonparametric maximum likelihood estimator and is the one that has to be trusted. The dotted line corresponds to the Aalen-Johansen estimator based on the interval midpoints. The difference is more expressed in the beginning of the study period. Finally the dashed line is obtained with the Kaplan-Meier estimator based on the midpoints. This estimate suffers both from the ad-hoc midpoint approach and from the apparently wrong treatment of the observations where exfoliation or another competing risk occurred before the fracture of the filling.

15.4 Regression models

15.4.1 Cox regression

Regression models describe the influence of certain factors on the event times (see also Chapter 11 for a general introduction into regression models). These factors are generally called covariates; they refer to characteristics and measurements

Table 15.1 Cox regression analysis of implant failure.

Implant specific factors.				
Factor	$\widehat{\beta}$	Hazard ratio	CI-95 %	p-value
upper vs lower jaw	1.27	3.56	1.22–10.41	0.02
implant length (mm)	0.040	1.04	0.72–1.50	0.83
Patient specific factors				
Factor	$\widehat{\beta}$	Hazard ratio	CI-95 %	p-value
female vs male	−0.32	0.724	0.25–2.11	0.55
age (years)	0.046	1.047	1.011–1.08	0.01

taken on the patient and the study unit. More specifically baseline covariates provide information that is available at the time origin and time-dependent covariates provide information that is updated during the study period. Due to censoring it is usually not possible to relate the event times or their expected values directly to the covariates. Instead most of the commonly used survival regression models specify the relation between the factors and the hazard rate.

Table 15.1 shows the results of a Cox regression analysis based on the TPS-SteriOss® implant data. Considered are the following covariates: patients age in years, gender, implant length in mm, and implant jaw position.

The hazard ratio is used for example to quantify how much more prone dental implantation is in the upper jaw is compared to the lower jaw; it is obtained as

$$\text{hazard ratio} = \frac{\text{hazard}(t \mid \text{upper jaw})}{\text{hazard}(t \mid \text{lower jaw})} = \exp(\hat{\beta}) = \exp(1.27) = 3.56.$$

Thus, the TPS-SteriOss® implant data show that the hazard for implant failure is about 3.5 times higher in the upper jaw. The 95 % confidence interval for the hazard ratio is 1.22–10.41 indicating the uncertainty about the true risk multiplier. To account for the fact that multiple implants are studied in the same mouth a modified variance formula was applied to obtain the confidence intervals [10].

The Cox regression model is also called proportional hazard model [23]. This model postulates that covariate effects do not depend on time. In particular, this implies that the survival functions corresponding to two different study units can not cross. The basic assumptions is that the hazard rate conditional on the event history (see Section 15.2.6) is on the form:

$$\lambda\left(t \mid Z_{ik}\right) = \lambda_0(t)\exp(\beta_1 Z_{ik}^{(1)} + \cdots + \beta_p Z_{ik}^{(p)})$$

given covariates $Z_{ik}^{(1)}, \ldots, Z_{ik}^{(p)}$ for the *'k'*th study unit of patient *'i'*. Here λ_0 is a function of time that describes the baseline hazard rate [23].

15.4.2 Time-varying effects

One limitation of the Cox proportional hazard model is that it cannot represent time-varying effects. Regarding the sometimes long follow-up periods in dental

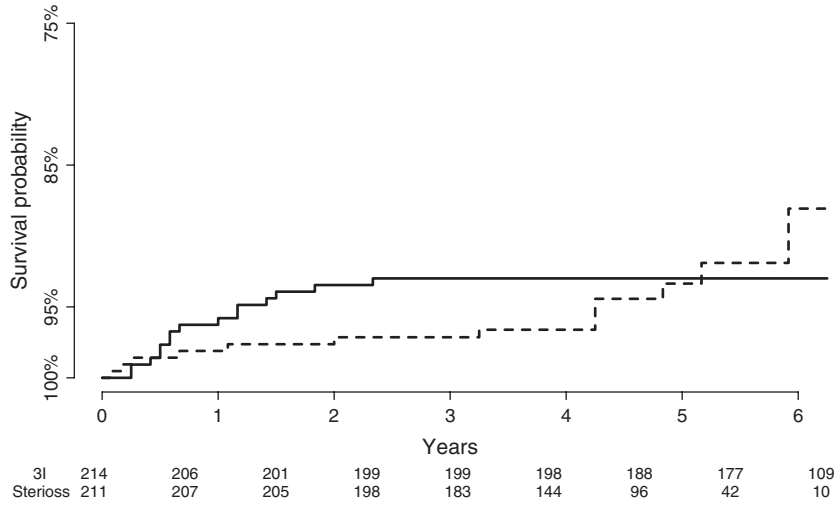


Figure 15.7 The survival functions for the TPS-SteriOss® implant system (dashed line) and the 3I® implant system (solid line) cross in time.

applications this assumption is critical. Figure 15.7 shows Kaplan-Meier curves for the TPS-SteriOss® [1] and the 3I® implant system [24]. The corresponding hazard rates are non proportional.

There are some alternative models that are well suited for exploring time-varying effects. One important alternative to the proportional hazards model is Aalen's additive hazard regression model [25; 26; 27] that will allow that all covariate effects vary with time, and where the form of the hazard conditional on the event history is

$$\lambda(t \mid Z_{ik}) = \beta_1(t)\, Z_{ik}^{(1)} + \cdots + \beta_p(t)\, Z_{ik}^{(p)}.$$

Here the covariate effects are interpreted as excess risk, i.e. the increase of risk relative to some baseline risk. In the additive model the hazard of a study unit is a sum of the effects of the covariates. In the example above that compares TPS-SteriOss® and 3I® implants, the additive model compares the survival of the implants without any parametric assumptions and would lead to an excess hazard for 3I® compared to TPS-SteriOss® that is simply the hazard for 3I® minus the hazard for TPS-SteriOss®. This effect of the implant type would be positive initially and negative later, thus reflecting the crossing survival curves in Figure 15.7.

From a practical perspective a useful sub-model is the semi-parametric additive risk model [28], where some of the effects are constant and the so-called **Cox-Aalen** model that assumes that the effects of some covariates are additive and time-varying while others act multiplicatively and time-constantly on the hazard rate [29; 30].

$$\lambda(t \mid Z_{ik}) = \left\{ \beta_1(t)\, Z_{ik}^{(1)} + \cdots + \beta_q(t)\, Z_{ik}^{(q)} \right\} \exp(\beta_{q+1}\, Z_{ik}^{(q+1)} + \cdots + \beta_p\, Z_{ik}^{(p)}).$$

Here the hazard has two components. The covariate effects in the first (additive) component has an excess risk interpretation as explained above, and the covariate effects in the second (multiplicative) component are interpreted as relative risks [31].

Consider a Cox proportional hazard regression model for the cause-specific hazard rate of all filling failures in the primary teeth data as in [6]. The cumulative residual test for proportionality [32] shows that the effects of yes or no endodontic treatment ($p = 0.034$) and children's age ($p = 0.009$) are significantly time-varying. The right panel of Figure 15.8 compares what would be expected if the proportional hazard assumption was correct (thin grey lines) to what has been observed (thick line) for the effect of endodontic treatment. If the Cox regression model was indeed the correct model the process that reflects the fit of the model (thick line) should vary as the simulations (thin grey lines). In the current example the observed process is not consistent with its behaviour under the Cox model and the proportional hazards assumption can be rejected for the factor endodontic treatment. A similar plot for age is not shown.

A Cox-Aalen model can be fitted using the R-add-on-package 'timereg' in which filling material, cavity type and treatment problems act multiplicatively and time-constant on the hazard rate but endodontic treatment and age are additive and time-varying. Even though the Cox model and the Cox-Aalen models are not formally nested, in practice the Cox-Aalen model provides an extension that allows

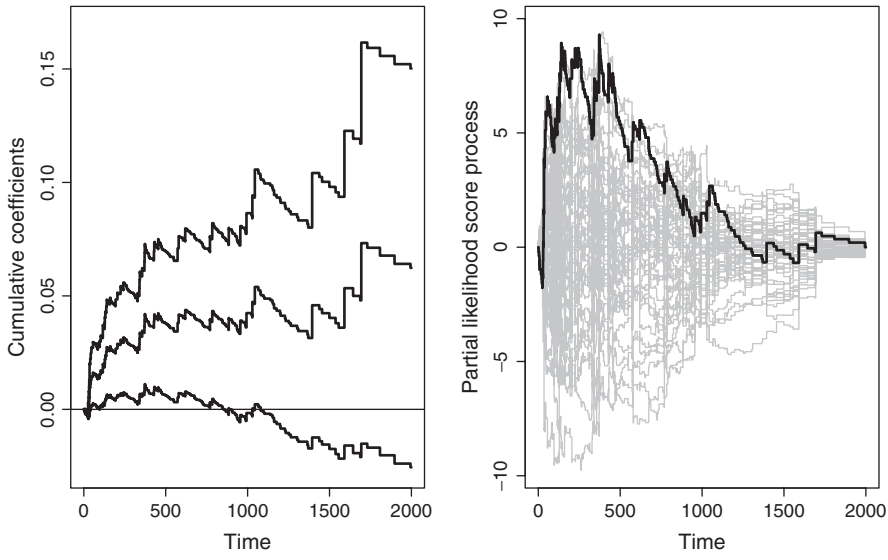


Figure 15.8 Left panel: Time-varying effect of endodontic treatment on the hazard rate of filling failure (left panel) with 95% pointwise confidence limits. Right panel: Observed score process (thick line) and 50 bootstrap simulations (thin grey lines) under the proportional hazards assumption of the Cox model.

endodontic treatment and age to have time-varying effects on the hazard rate. Note that the interpretation is that these covariates lead to excess hazard rates rather than relative risk. The left panel of Figure 15.8 shows the time-varying regression coefficients of endodontic treatment (the curve in the middle) and confidence bands. The slope of the cumulative excess hazard rate for endodontic treatment thus reflects the excess hazard rate corrected for the other covariates. Note that the excess risk is increased by about 14/10000 in the first 500 days (by reading of the slope of the cumulative for the first 500 days) and then essentially disappears with an approximate excess rate of less than 3/10000. Note also that the early excess risk has a large early component. A standard Cox regression model where all covariate effects are proportional, in contrast, would incorrectly say that endodontic treatment leads to a relative risk of 1.733 (95 % confidence interval: (1.258, 2.387)) over the entire time-span of the study. In this example the effect sizes of the restorative material, treatment problems, cavity type did not change very much by introducing the time-varying effects when compared to the original analysis.

15.4.3 Frailty models

It may often be natural to assume that events in the same cluster are correlated, where examples of relevant cluster structures are several fillings in the same tooth, several teeth in the same mouth, several patients treated by the same dentist. A simple representation of this phenomenon is given by frailty models [33] which assume that all hazard rates in a cluster are multiplied by the same factor, called the frailty variable, and that this frailty variable varies across clusters according to some distribution. Note that a frailty variable is similar to a random intercept in a random effects model introduced in Chapter 13. The gamma distribution is mathematically convenient and a popular choice.

The Cox regression model with frailty term is

$$\lambda_{ik}(t) = \lambda_0(t)\,\Gamma_k\,\exp(\beta_1 Z_{ik}^{(1)} + \cdots + \beta_p Z_{ik}^{(p)}).$$

For the TPS-SteriOss® implant data a Cox regression model with a gamma frailty term yields that the rate ratio between implant survival in the lower jaw compared to the upper jaw is about. In contrast, the same rate ratio estimated from the model without frailty term is 3.5, see Section 15.4.1. This attenuation of effect in the model without frailty term is a typical finding which can be explained as follows.

A useful way of thinking about clustered data is that there is an important but apparently unobserved factor which would explain the dependence structure of the data. For example there could be a gene which make some patients more susceptible to inflammation and hence to loosing an implant. To include a frailty variable in a hazard model is an attempt to compensate for the missing information of the unobserved important factor. For the Cox proportional hazard model it is known that excluding an important factor from the analysis leads to estimates of the covariate effects which are systematically too small in absolute value, i.e. closer

to zero than the true effects [34]. Thus, effect estimates obtained with a frailty Cox regression model will often yield higher effect estimates as compared to those obtained with a corresponding Cox regression model without frailty variable.

As explained, the aim is that the frailty variable Γ_k reflects unobserved heterogeneity between clusters and boosts or lowers the failure rates jointly for all study units within the same cluster. In addition to the covariate effects one can be interested in the magnitude of heterogeneity that is not explained by the observed covariates. This can be quantified by the variance of the frailty variable. To estimate this variance a maximum likelihood [35] or a penalized likelihood approach [36] can be used.

In the dental implant example, the gamma frailty variance is estimated as 4.4. To interpret this quantity one can use Kendall's coefficient of concordance which can directly be calculated from the frailty variance. Kendall's coefficient of concordance (plus constant 0.5) is the probability that two identically distributed pairs of failure times from different clusters are concordant. For the dental implant example Kendall's coefficient of concordance is 0.7 implying high concordance.

15.5 Remarks

In this chapter we have focused on nonparametric and semiparametric methods mainly to explain how the information of the censored data enters the statistical analysis. However, once the type of censoring and competing risks have been recognized then other methods can provide more stable and feasible estimates in particular for small and moderate sample sizes.

Parametric survival models can be fitted using standard software also when the event times are interval censored [17]. For example, [37] considered the accelerated failure time model in dental applications. However, parametric models often make strong assumptions. It is therefore important that the results are in agreement with a corresponding (nonparametric) approach which does not make these assumptions. An exemplary investigation where parametric models are justified in that way is [38].

The Bayesian approach has been successfully applied in dental applications with complex observational patterns [39; 40; 41]. In terms of practical ease the class of discrete time survival regression models is also highly attractive [42]. Various functions are possible to link the covariates to the event times and random effects (frailty variables) can be added even separately for multiple cluster levels. These models are fitted using standard software for generalized linear models [43]. The extensions to competing risks are readily available [44].

As mentioned in the introduction the aim is often to guide dental decision making by using predictions. The estimates of Section 15.3 can usually be directly interpreted as predictions. But using the information of patient and study unit specific factors usually increases the predictive accuracy. References 45; 46 discuss this in the context of predicting survival of dental implants.

References

[1] T. A. Gerds & M. Vogeler (2005) Endpoints and survival analysis for successful osseointegration of dental implants. *Statistical Methods in Medical Research*, **14**(6), 579–90.

[2] E. Marubini & M. G. Valsecchi (1995) *Analyzing Survival Data from Clinical Trials and Observational Studies*. Statistics in Practice. Chichester: John Wiley & Sons, Ltd.

[3] J. D. Kalbfleisch & R. L. Prentice (2002) *The Statistical Analysis of Failure Time Data*. 2nd ed. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons, Ltd.

[4] J. Grüger, R. Kay & M. Schumacher (1991) The validity of inference based on incomplete observations in discrete state models. *Biometrics*, **47**(2), 595–605.

[5] P. K. Andersen, Ø. Borgan, R. D. Gill & N. Keiding (1993) *Statistical Models Based on Counting Processes*. Springer Series in Statistics. New York, Springer.

[6] V. Qvist, L. Laurberg, A. Poulsen & P. T. Teglers (2004) Eight-year study on conventional glass ionomer and amalgam restorations in primary teeth. *Acta Odontologica Scandinavia*, **62**(1), 37–45.

[7] J. Lindsey & L. Ryan (1998) Tutorial in biostatistics - methods for interval-censored data. *Statistics in Medicine*, **17**(2), 219–38.

[8] N. Keiding, K. Begtrup, T. H. Scheike & G. Hasibeder (1996) Estimation from current-status data in continuous time. *Lifetime Data Analysis*, **2**(2), 119–29.

[9] T. R. Walton (1998) The outcome of implant-supported fixed prostheses from the prosthodontic perspective: proposal for a classification protocol. *International Journal of Prosthodontics*, **11**(6), 595–601.

[10] S. Chuang, L. Wei, C. Douglass & T. Dodson (2002) Risk factors for dental implant failure: A strategy for the analysis of clustered failure-time observations. *Journal of Dental Research*, **81**(8), 572–7.

[11] R. L. Williams (1995) Product-limit survival functions with correlated survival times. *Lifetime Data Analysis*, **1**(2), 171–86.

[12] Z. Ying & L. Wei (1994) The Kaplan-Meier estimate for dependent failure time observations. *Journal of Multivariate Analysis*, **50**(1), 17–29.

[13] S. Chuang, L. Tian, L. Wei & T. Dodson (2001) Kaplan-Meier analysis of dental implant survival: A strategy for estimating survival with clustered observations. *Journal of Dental Research*, **80**(11), 2016–20.

[14] R. L. Williams (2000) A note on robust variance estimation for cluster-correlated data. *Biometrics*, **56**(2), 645–6.

[15] B. Efron & R. J. Tibshirani (1993) *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. New York: Chapman & Hall.

[16] P. M. Odell, K. M. Anderson & R. B. D'Agostino (1992) Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, **48**(3), 951–9.

[17] J. P. Klein & M. L. Moeschberger (2003) *Survival Analysis. Techniques for Censored and Truncated Data*. 2nd ed. Statistics for Biology and Health. New York, Springer.

[18] A. Shick & Q. Yu (2000) Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics*, **27**(1), 45–55.

[19] R. Peto (1973) Experimental survival curves for interval-censored data. *Applied Statistics*, **22**, 86–91.

[20] B. Turnbull (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society*, Series B, **38**, 290–5.

[21] J. A. Wellner and Y. Zhan (1997) A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association*, **92**(439), 945–59.

[22] M. G. Hudgens, G. A. Satten & I. M. Longini (2001) Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics*, **57**(4), 74–80.

[23] D. R. Cox (1972) Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*. Series B, **34**, 187–220.

[24] M. Knauf, T. Gerds, R. Muche & J. R. Strub (2007) Survival and success rates of 3i implants in partially edentulous patients: results of a prospective study with up to 84-months' follow-up. *Quintessence International*, **38**(8), 643–51.

[25] O. Aalen, Ø. Borgan, N. Keiding & J. Thormann (1980) Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics*, **7**, 161–71.

[26] O. O. Aalen (1989) A linear regression model for the analysis of life times. *Statistics in Medicine*, **8**(8), 907–25.

[27] O. O. Aalen (1993) Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine*, **12**, 1569–88.

[28] I. W. McKeague & P. D. Sasieni (1994) A partly parametric additive risk model. *Biometrika*, **81**(3), 501–14.

[29] T. H. Scheike & M.-J. Zhang (2002) An additive-multiplicative Cox-Aalen model. *Scandinavian Journal of Statistics*, **29**(1), 75–88.

[30] T. H. Scheike & M.-J. Zhang (2003) Extensions and applications of the Cox-Aalen survival model. *Biometrics*, **59**(4), 1033–45.

[31] T. Martinussen & T. Scheike (2006) *Dynamic Regression Models for Survival Data*. Statistics for Biology and Health. New York, Springer.

[32] D. Lin, L. Wei & Z. Ying (1993) Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, **80**(3), 557–72.

[33] P. Hougaard (2000) *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. New York, Springer.

[34] C. A. Struthers & J. D. Kalbfleisch (1986) Misspecified proportional hazard models. *Biometrika*, **73**, 363–9.

[35] G. G. Nielsen, R. D. Gill, P. K. Andersen & T. I. A. Sørensen (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, **19**(1), 25–43.

[36] T. M. Therneau & P. M. Grambsch (2000) *Modeling Survival Data*. Statistics for Biology and Health. New York, Springer.

[37] E. Lesaffre, A. Komarek & D. Declerk (2005) An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, **14**(6), 539–52.

[38] O. O. Aalen, E. Bjertness & T. Sonju (1995) Analysis of dependent survival data applied to lifetimes of amalgam fillings. *Statistics in Medicine*, **14**, 1819–29.

[39] M. C. M. Wong, K. F. Lam & E. C. M. Lo (2005) Bayesian analysis of clustered interval-censored data. *Journal of Dental Research*, **84**(9), 817–21.

[40] T. Härkänen, M. A. Larmas, J. I. Virtanen & E. Arjas (2002) Applying modern survival analysis methods to longitudinal dental caries studies. *Journal of Dental Research*, **81**(2), 144–8.

[41] A. Komárek, E. Lesaffre, T. Härkänen, D. Declerck & J. I. Virtanen (2005) A Bayesian analysis of multivariate doubly-interval-censored dental data. *Biostatistics*, **6**(1), 145–55.

[42] L. Fahrmeir (2005) Discrete failure time models. In *Encyclopedia of Biostatistics*, volume **2**, pp. 1458–63. John Wiley & Sons, Inc., New York.

[43] L. Fahrmeir & G. Tutz (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2nd ed. Springer Series in Statistics. New York, Springer.

[44] L. Fahrmeir & S. Wagenpfeil (1996) Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association*, **91**(436), 1584–94.

[45] S.-K. Chuang & T. Cai (2006) Predicting clustered dental implant survival using frailty methods. *Journal of Dental Research*, **85**(12), 1147–51.

[46] S. Chuang, L. Tian, L. Wei & T. Dodson (2002) Predicting dental implant survival by use of the marginal approach of the semi-parametric survival methods for clustered observations. *Journal of Dental Research*, **81**(12), 851–5.

16

Misclassification and measurement error in oral health

Helmut Küchenhoff

16.1 Introduction: measurement problems in oral health

When doing statistical inference, we usually assume that the data being used are correctly measured. However, in many studies the variables of interest cannot be observed directly or measured correctly. One example in oral health is the measurement of caries. In a recent paper Ismail (2004) states 'In spite of the progress in understanding the caries process, there is still some significant level of confusion among members of the dental community on what is dental caries.' This statement says that there is not only a problem of measuring caries but also a problem of defining caries.

Another important example for measurement problems in oral health research is assessing nutritional behavior of study participants. In a recent study Dion *et al*. (2007) show that the relationship between malnutrition and oral health is estimated with bias, when there is error in the measurement of malnutrition. Again there is a serious problem in defining and measuring malnutrition.

The problems of data validity and quality have been discussed in Chapter 9 of this book in detail. Furthermore, reliability of diagnostic tests was the issue in Chapter 12. In those chapters the focus was on assuring data quality, avoid

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

measurement error and quantify the problems in diagnosis by estimating sensitivity and specificity of a diagnostic tool. It is common practice, that assuring data quality is a preprocessing step apart from the main data analysis. After minimizing measurement error, the analysis of the data is performed assuming that there is no measurement error. This can be adequate in many practical situations, but it can lead to bias, lack of power in statistical tests and wrong conclusions, if measurement error is still inevitably high.

Let us consider as an example prevalence of caries. It can be estimated based on a random sample by the proportion of children, for which the dental examiner diagnosed caries. For illustration, we assume that caries is overlooked by the examiner in e. g. 50 % of the diseased children and there are no false positive diagnosis (i.e. the sensitivity of the diagnosis is 0.5 and the specificity is 1). In principle, one has two options to cope with this problem. First, one can try to improve scoring by training or giving more time for examination. Second, one can correct the study results. If the observed prevalence is 8 % then a corrected estimation would obviously be 16 %, because half of the caries cases are missed by the examiner. Thus, underscoring of the examiner is taken into account when presenting the results. In practice, things are not as easy. In the Signal-Tandmobiel® study (Vanobbergen *et al*. (2000), see Chapter 19 of this book) many examiners were involved and it turned out that the underscoring of the examiner differed considerably in different regions. The question arises whether observed differences between regions are artifacts induced by the different scoring behavior of the examiners. This example is discussed further at the end of this chapter.

We present methods for connecting considerations about data quality to the main statistical analysis of a study. The key idea is to introduce a statistical model for the measurement process, which is called the *measurement model* and include this into the data analysis.

16.1.1 Basics

Going back to the problems of defining caries the question of an oral health study ‘What are the relevant predictors for caries’ or ‘Is there a causal relationship between caries and sugar intake’ is meaningless when there is no clear definition of ‘caries’. As a first essential step, one has to make precise definitions of the variables to be included. If there is no clear substantial definition like e.g. for the body mass index, then operational definitions of the variables have to be used. Typical examples are Intelligence quotient, scores from questionnaires or other scores like the Gingival index in oral health. Sometimes one has to go a further step by using a benchmark measurement instead of an operational definition. For example, one method or one examiner can be seen as the best possibility for measuring caries. The definition of the variables used in the study should be taken into account when the results are interpreted. For a general more philosophical view on measuring problems, see Hand (2004).

After defining the correct variable, which is also called *gold standard*, in a second step the question of correct measurement has to be addressed. In assessing

caries, the diagnosis of a dental examiner can be different from the correct diagnosis. This phenomenon is called *misclassification* for a binary or discrete variable. If the variable of interest is continuous then the term *measurement error* is commonly used. Because in both cases there are measurement problems, we use *measurement error* as a general term.

16.1.2 Notation and Literature

For regression problems, we use the standard notation with the response variable Y and the predictors denoted by X and if needed Z. Unfortunately, there is no standard notation in the literature for the variables with possible measurement error. In the following we use the '*' - notation, i.e. when Y or X is the variable of interest, then we denote the observed and possibly corrupted variable by Y^* and X^*, respectively.

There is a rich literature on measurement error problems, including two recent books. The book by Carroll *et al*. (2006) gives an excellent overview of the current state of the art, while the book by Gustafson (2004) has a focus on misclassification problems and takes a Bayesian perspective on the issue. Both books are written on a technical level, so that one should inspect Kuha *et al*. (2001) for an overview.

16.2 Misclassification

16.2.1 Prevalence estimation

Going back to the example mentioned above we are interested in the prevalence of caries, more specifically the proportion of children with caries experience (CE), i.e. children with at least one of the deciduous teeth is decayed, missing or filled due to caries. Suppose in a population 2000 of 10,000 six-year-old children have CE, then the prevalence is 0.2. Let us denote the random variable, which describes the true status of a child i by Y_i with $Y_i = 1$ if child i has CE and $Y_i = 0$ otherwise. The relating measurement denoted by Y^* is e. g. the result of the dental examiner ($Y_i^* = 1$ for a diagnosis of CE and $Y^* = 0$ for the diagnosis of no CE). Because it is not known whether one particular observation is misclassified or not, we use a stochastic model for the misclassification, i.e. we describe our lack of information by probabilities. This model is called the *measurement model*. We distinguish this model from the model of primary interest, for which we use the term *main model*. In our case of a binary variable with possible outcomes 0 and 1, the measurement model can be described by $\pi_{11} = P(Y^* = 1|Y = 1)$, which is the probability of observing CE, when the child has CE, $\pi_{00} = P(Y^* = 0|Y = 0)$, which is the probability of observing no CE if the child has no CE. In the example of a diagnostic test, the π_{11} is known as sensitivity or true positive fraction (TPF) and π_{00} as specificity or 1- true negative fraction (1- TNF), see also Chapter 12. The measurement model has the simple form

$$P(Y^* = 1|Y = 1) = \pi_{11} \quad \text{(sensitivity)},$$

$$P(Y^* = 0|Y = 0) = \pi_{00} \quad \text{(specificity)},$$

$$P(Y^* = 0|Y = 1) = 1 - \pi_{11} = \pi_{01},$$
$$P(Y^* = 1|Y = 0) = 1 - \pi_{00} = \pi_{10}.$$

We assume that we have a simple random sample of size n from the population, i.e. we observe the results from the dental examiner, which are denoted by $Y_1^*, \ldots, Y_n^*$. When we ignore possible misclassification the prevalence p is estimated by the relative frequency of children diagnosed with caries:

$$\hat{p}_{na} = \frac{1}{n} \sum_{i=1}^{n} Y_i^*. \tag{16.1}$$

As it is well established in the measurement error literature, we call the estimator that simply ignores measurement error the *naive estimator*. The expected value of the prevalence estimation is given by

$$E(\hat{p}_{na}) = P(Y^* = 1) = \pi_{11} P(Y = 1) + \pi_{10} P(Y = 0). \tag{16.2}$$

Formula (16.2) takes into account that a diagnosis of caries can be correct ($Y = 1$) (true positive) or incorrect ($Y = 0$) (false positive). Going back to the example of the hypothetical population of 10,000 children, of which 2000 have CE, we assume that the sensitivity is 0.7. Then, on average 0.7*2000 = 1400 will be correctly scored having CE. If the specificity is 0.9 then the false positive rate is 0.1 and on average 0.1*8000 = 800 children will be scored false positive. The prevalence will be estimated by 1400+800/10000 = 0.22, which differs from the true value 0.2. In general, there will be a difference between the expected naive prevalence estimation $E(\hat{p}_{na})$ and the true prevalence. This difference is called *bias* and it depends on the true prevalence and the misclassification matrix. In Table 16.1 some calculations for different scenarios of prevalence, sensitivity and specificity are performed. It can be seen that the bias can go in either direction.

Equation (16.2) gives an explicit relationship between the observed prevalence $P(Y^* = 1)$ and the true prevalence $P(Y = 1)$. A simple algebraic transformation yields

$$P(Y = 1) = \left[P(Y^* = 1) - \pi_{10}\right] / (\pi_{11} + \pi_{00} - 1). \tag{16.3}$$

Table 16.1 Calculations for different values of prevalence, specificity and sensitivity.

True prevalence $P(Y = 1)$	Specificity π_{00}	Sensitivity π_{11}	Observed prevalence $P(Y^* = 1)$	Bias
0.5	0.8	0.8	0.5	0
0.9	0.9	0.9	0.82	−0.08
0.8	0.9	0.99	0.81	0.01

If we assume that the misclassification probabilities are known, the true prevalence can be estimated without bias and its variance can be calculated as follows:

$$\hat{p} = \left(\frac{1}{n} \sum_{i=1}^{n} Y_i^* - \pi_{10} \right) / (\pi_{11} + \pi_{00} - 1), \tag{16.4}$$

$$Var(\hat{p}) = Var \left(\frac{1}{n} \sum_{i=1}^{n} Y_i^* \right) / (\pi_{11} + \pi_{00} - 1)^2. \tag{16.5}$$

Equation (16.4) is the result of one basic strategy for measurement error correction: Find the distribution of the observed variable (here the observe prevalence $P(Y^* = 1)$) and use the measurement model for estimating the parameters of interest (here the true prevalence $P(Y = 1)$). Note that the variance of $\hat{p}$ is higher than the variance without misclassification resulting in wider confidence intervals. Simply ignoring misclassification gives a precise but biased (i.e. possibly wrong) result while taking the uncertainty of data into account leads to wider, but correct confidence intervals. Furthermore, we have to assume that the sum of specificity and sensitivity is greater than 1, i.e. $\pi_{00} + \pi_{11} - 1 > 0$. This is a sensible assumption, because otherwise the classification process has basically no information.

16.2.2 Estimation of misclassification probabilities

The main practical problem in applying correction formula (16.4) is that sensitivity and specificity have to be known. Otherwise no correction is possible. This a typical situation for measurement error modeling. One needs some information about the parameters of the measurement model to obtain a corrected estimation in the main model. In principle, there are three possibilities for gaining this information, which are briefly discussed in following.

16.2.2.1 Making assumptions (external validation)

If no direct information is available, then one can use results from other studies with a similar measuring process. One has to be very careful, because the transferability of the misclassification probabilities to the current study has to be assured. Using external validation studies has the character of making assumptions. Note that simply ignoring misclassification is making the assumption that the misclassification probabilities are 0, which can be as problematic as using external information. Another possibility is to use different scenarios for the misclassification probabilities and interpret the results as a sensitivity analysis.

16.2.2.2 Internal validation with gold standard

The best possibility to estimate the relationship between the correct variable Y and the corresponding measurements Y^* is to perform an internal validation study, where both Y and Y^* are available. Ideally one uses a random sample from the

study population of interest and uses the standard measurement method of the main study producing Y^*. Then, by more sophisticated measurement methods the true (gold standard) variable Y is sampled. This can be a gold standard examiner or an X−ray in caries research. Then, sensitivity and specificity can be estimated as relative frequencies from the validation data. In addition, more complex modeling is possible, if the information of the validation study contains more information. Lesaffre *et al*. (2009) propose a misclassification model using single tooth information in the validation study.

16.2.2.3 Validation with replicate measurements on the same subject

When no gold standard is available, then one could use replicate measures on the same subjects. This is a good method for continuous measurement error but has some problems for the misclassification case. The main problem is that sensitivity and specificity cannot be estimated from repeated measures. Suppose we have two observations Y_{i1}^* and Y_{i2}^* for each subject in the validation data. We assume independence of the replicates and identical misclassification probabilities given the true value Y_i. These assumptions yield the following probabilities.

$$\begin{aligned} P(Y_{i1}^* = 1, Y_{i2}^* = 1) &= p\pi_{11}^2 + (1-p)\pi_{10}^2, \\ P(Y_{i1}^* = 1, Y_{i2}^* = 0) = P(Y_{i1}^* = 0, Y_{i2}^* = 1) &= p\pi_{11}\pi_{01} + (1-p)\pi_{10}\pi_{00}, \\ P(Y_{i1}^* = 0, Y_{i2}^* = 0) &= p\pi_{01}^2 + (1-p)\pi_{00}^2. \end{aligned}$$

Because the last equation is redundant (sum of probabilities $= 1$), this is a system of two equations for the three unknown parameters p, π_{00}, π_{11}. The parameters cannot be estimated from the joint distribution of (Y_{i1}, Y_{i2}). The parameters can be estimated, when there are more than two measurements on one subject. Then, methods of latent class analysis can be applied, for a recent discussion see Pepe and Janes (2007). A further practical problem is the independence assumption of the replicates. When in the diagnosis of caries three different examiners assess caries at the same child, there could be induced some independence because there are children, which may be easily diagnosed and others are not. Because of that problem, in most studies reliability measures are calculated instead of estimating sensitivity and specificity. The most common approach is to calculate the inter rater reliability coefficient κ, see Chapter 9. It indicates the agreement between two raters, but cannot be used for correction like in Equation (16.4). For illustration, in Figure 16.1 possible values for sensitivity and specificity when $\kappa = 0.5$ and the true prevalence $p = 0.2$ is displayed.

Note that the specificity ranges between 0.54 and 1 while the sensitivity is between 0.85 and 1. Using (16.2) the expected value of the naive estimator ranges from 0.17 to 0.57. We conclude that a correction for misclassification is usually not possible. For higher values of κ, a worst case scenario analysis as a sensitivity analysis is possible. In the presence of high misclassification rates, a validation study using a gold standard is desirable.

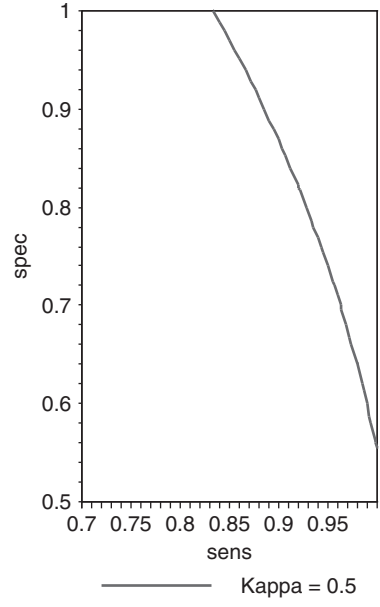


Figure 16.1 Possible values of sensitivity and specificity for $\kappa = 0.5$ *and* $p = 0.2$.

16.2.3 Combining measurement model and main model

After estimating sensitivity and specificity, this uncertainty has to be taken into account for the variance estimation of the prevalence estimator $\hat{p}$ defined by (16.4). Denoting the variance estimators from the validation study by $Var(\hat{\pi}_{00})$ and $Var(\hat{\pi}_{11})$ an estimator for $Var(\hat{p})$, is as follows, see Greenland (1988):

$$Var(\hat{p}) = \frac{\hat{p}^*(1-\hat{p}^*)}{(\pi_{00}+\pi_{11}-1)^2} + Var(\hat{\pi}_{00})\frac{(\pi_{11}-\hat{p}^*)^2}{(\pi_{00}+\pi_{11}-1)^2}$$
$$+Var(\hat{\pi}_{11})\frac{(1-\pi_{00}-\hat{p}^*)^2}{(\pi_{00}+\pi_{11}-1)^2}. \qquad (16.6)$$

The following example shows the consequences of this rather simple calculations. In the Signal-Tandmobiel® study (Vanobbergen *et al.* (2000)), 4468 children in Flanders were examined in the years between 1996 and 2001. For illustration, we ignore the longitudinal structure of the data and discuss prevalence estimation of caries in the first 4 molars. Because the children were about 7 years old at study onset this can be seen as a prevalence estimation for 7-year-old children in 1996, for 8-year-old children in 1997 etc. A validation study has been performed in the years 1996, 1998, 2000. A gold standard examiner was compared to the examiners of the study ($n_{1996} = 142, n_{1998} = 157, n_{2000} = 148$). In Table 16.2, the results for the prevalence estimation are summarized. Note that there is substantial misclassification in the data and that the 95 % confidence intervals are much wider than the confidence intervals ignoring misclassification.

Table 16.2 Results for prevalence estimation in the Signal-Tandmobiel® study. The naive prevalence estimation, the naive Confidence interval (CI), specificity estimation (Spec), sensitivity estimation (Sens), the corrected estimation (16.4) and the confidence interval based on (16.6) are reported.

Year	Naive prev	Naive CI	Spec	Sens	Corr. prev	Corr. CI
1996	0.10	0.095; 0.114	0.93	0.61	0.06	0; 0.15
1997	0.20	0.185; 0.210	0.93	0.61	0.23	0.13; 0.34
1998	0.28	0.265; 0.294	0.87	0.73	0.25	0.15; 0.35
1999	0.34	0.325; 0.356	0.87	0.73	0.35	0.25; 0.45
2000	0.38	0.364; 0.397	0.86	0.93	0.31	0.24; 0.37
2001	0.42	0.398; 0.432	0.86	0.93	0.35	0.28; 0.42

16.2.4 Misclassification in more complex discrete data: the matrix method

The basic ideas of the prevalence estimation can be expanded to more complex discrete data situations. Suppose we want to compare two probabilities by a 2×2 table, e.g. prevalence of caries in boys and girls or compare two groups in a clinical trial, then equation (16.2) can be applied for each group. Then, one has to solve the equation for the two true probabilities of interest, see (16.3). This yields e. g. a corrected estimator for an odds ratio or for a risk difference. In a more general setting (e.g. more complex contingency tables) a system of linear equations has to be solved, which can be done by matrix calculations. Again, information about the misclassification probabilities is needed. The general procedure is called the matrix method, for more details see Kuha *et al*. (2001) and Morrissey and Spiegelman (1999).

16.2.5 Logistic regression with misclassified outcome

Assume we are interested in the relationship of caries to predictors like brushing behavior, age, gender etc. Then, usually a logistic model is applied:

$$P(Y = 1|X) = L(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k).$$

Here Y is the indicator for caries, x_j are the predictor variables and L the logistic distribution function $L(t) = 1/(1 + exp(-t))$. We observe the possibly incorrect diagnosis of the examiner, denoted by Y^*. Like in prevalence estimation, see (16.2), we can also write down the relationship between the predictors and the diagnosis of the examiner. For notational convenience, we define the linear predictor as follows.

$$h(x, \boldsymbol{\beta}) := \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

The observed data are described by

$$P(Y^* = 1|X) = \pi_{11} L\left[h(x, \boldsymbol{\beta})\right] + (1 - \pi_{10})\left\{1 - L\left[h(x, \boldsymbol{\beta})\right]\right\}.$$

The parameter of interest β_k can now be estimated from our observed data, if sensitivity and specificity are known or can be estimated from validation data. For details and further developments, see Neuhaus (1999).

16.3 Measurement error in continuous variables

16.3.1 Basic measurement error models

16.3.1.1 Classical measurement error

As an example, we consider the variable fluoride concentration in the public water supply of a community. Since the fluoride concentration may vary over time, the gold standard variable X is defined as a long term average fluoride concentration. When one single measurement is taken, then the measured value X^* differs form the true value X. For illustration, w e use a small simulated data set of 10 observations, see Table 16.3. A possible measurement model is

$$X^* = X + U. \tag{16.7}$$

Here, U is the measurement error and it is assumed that $E(U) = 0$, $Var(U) = \sigma_U^2$ and U is independent of the true variable X. X^* differs randomly from the gold standard due to the time of the day when the measurement is taken and possibly due to a measurement error of the technical device. Note that the assumption $E(U) = 0$ means that there is no systematic part in the error. Due to independence, the variance of X^* is given by

$$Var(X^*) = Var(X) + \sigma_U^2.$$

Measurement error causes some extra variability to the data. The term

$$r = \frac{Var(X)}{Var(X^*)} = \frac{Var(X)}{Var(X) + \sigma_U^2}. \tag{16.8}$$

is called the reliability ratio. It is the part of the variance of the observed data that can be attributed to the true variable. In our example data the variance of X^* is 0.90, which is greater that the variance of X (0.47). The reliability ratio is 0.47/0.90 $= 0.52$.

16.3.1.2 Assessing measurement error variance

As will be explained below one needs information about the amount of measurement error, i.e. the measurement error variance for using correcting methods. As presented in Section 16.2.2 there are three principal methods: (1) *Making assumptions* about the measurement error variance can be possible, if measurement error is related to a technical device, when its precision is known. (2) *Internal validation* with a gold standard, i.e. a random sample of (X_i^*, X_i), gives a sample of measurement errors $U_i = X_i^* - X_i$. The sample variance of the U_i is an estimator for σ_U^2.

Table 16.3 Simulated data: true and observed levels of fluoride concentration in 10 communities.

Community	Gold standard X	Observed value X^*	Measurement error U
1	0.50	0.22	−0.28
2	0.60	0.41	−0.19
3	0.63	0.41	−0.22
4	0.72	0.57	−0.15
5	0.83	1.27	0.44
6	1.20	0.85	−0.35
7	1.50	1.66	0.16
8	1.70	2.17	0.47
9	2.00	2.28	0.28
10	2.50	2.96	0.46

In the numerical example of Table 16.3, the variance is estimated being 0.08. (3) If there are two independent *replicate measurements* X^*_{i1} and X^*_{i2} in a validation study, we are able to give an estimation for the measurement error variance without knowing the gold standard. Because $Var(X^*_{i1} - X^*_{i2}) = 2 * \sigma^2_U$, the measurement error variance can be estimated by half of the variance of the replicate differences. Note that the independence of the two measurements given the true value X_i is essential to perform this method.

16.3.1.3 Berkson measurement error

Suppose we want to measure the fluoride concentration on an individual level, but there are no individual data available. Instead of the individual data the exposure on the community level is used. We denote the community level exposure by Z^* and the gold standard individual exposure by Z. Then, a measurement model is now given by

$$Z = Z^* + U. \tag{16.9}$$

The individual exposure Z differs from the area mean due to individual conditions. Here, U is independent of Z^* and not of the true value. The main difference to the additive measurement error model is that the variance of the true variable Z is higher than the variance of Z^*, because

$$Var(Z) = Var(Z^*) + Var(U).$$

The effect of the Berkson error is fundamentally different to that of an additive measurement error, see below. In the literature there are many examples for more complex measurement models, like multiplicative measurement error, combination of Berkson and classical measurement error etc., see e.g. Carroll *et al.* (2006).

16.3.2 Measurement error in the outcome, ANOVA and t-test

We start with the case of simple linear regression with measurement error in the outcome variable. The linear regression model is given by

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{16.10}$$

We assume an additive measurement error model, i.e. $Y^* = Y + U$ resulting in

$$Y^* = \beta_0 + \beta_1 X + \epsilon + U. \tag{16.11}$$

Because we assumed that U is independent of X and ϵ, we find that (16.11) is a linear regression model with an extra error. The performance of the naive analysis, i.e. simply ignoring the measurement error is a valid strategy for data analysis. The same argument applies for group comparisons (t−test and ANOVA). The measurement error adds on the sampling error when estimating a group mean. This enlarges the internal variance in ANOVA. Although the power of the t-test and of the F−test in ANOVA is reduced by measurement error, the results of the analysis ignoring measurement error are correct. Hence, extra measurement error modeling is not necessary for t−test and ANOVA.

16.3.3 Linear regression with Berkson error in the predictor

The effect of a Berkson error in the predictor is similar to the case of an additive error in the outcome variable. Assuming a simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{16.12}$$

and a Berkson type error, i.e. $X = X^* + U$, where U is independent of X^*, then

$$Y = \beta_0 + \beta_1 X^* + \beta_1 U + \epsilon. \tag{16.13}$$

Because U is independent of X^*, the relationship between the observed variables Y and X^* is a linear regression with the same slope β_1 as in the model of interest. The effect of measurement error is a higher variance of the regression error term. This causes lower precision of the estimates and power reduction for the testing the hypothesis $\beta_1 = 0$. However, inference form the naive model, i.e. simply using X^* instead of X is unbiased. This result is also valid for the multiple linear regression model, but does not apply for non−linear models. However, in many cases like in logistic regression the bias induced by a Berkson error is rather low.

16.3.4 Linear regression with classical error in the predictor

We consider the relationship between fluoride concentration in the community water supply and the average number of decayed, filled or missing teeth per child

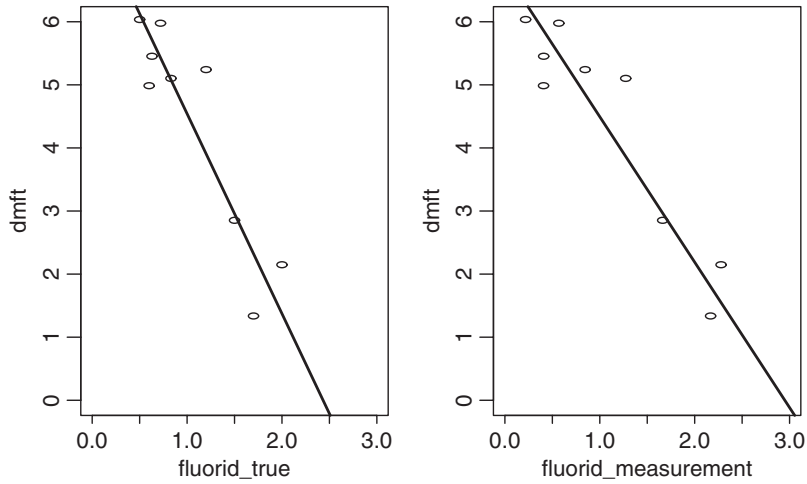


Figure 16.2 Effect of additive measurement error in linear regression. Example data with correct fluoride data (left panel) and fluoride data with additive classical measurement error (right panel).

(see Chapter 11). We use the example data from Table 16.3. In Figure 16.2, the scatter plots are displayed for the example data without measurement error (left panel) and for the data with measurement error (right panel). It can be seen that the slope of the regression line is more flat for the data with measurement error. One intuition is that the data are broadened in the x-direction by the measurement error. Another more general argument is that the relationship between the fluoride concentration and the outcome is partly hidden by the measurement error. We give now a more formal derivation of this effect, which is known as attenuation in the measurement error literature. The simple linear regression model is given by

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{16.14}$$

We assume an additive measurement error model, i.e. $X^* = X + U$. If we assume that X has a normal distribution $X \sim N(\mu_X, \sigma_X^2)$ and U is also normal with mean zero and variance σ_U^2, then we can calculate the observed model for the relationship of X^* and Y by

$$E(Y|X^*) = \beta_0 + \beta_1 E(X|X^*) = \beta_0 + \beta_1 (X^* - \mu_X) \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}.$$

The observed model is a linear regression with slope parameter $\beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}$, i.e. the attenuation factor is given by the reliability, see (16.8). Therefore, the naive (ordinary least squares) estimator $\hat{\beta}_{na}$ based on the observations (Y_i, X_i^*) is biased. Similar to the misclassification case the naive estimator can be corrected, when the

measurement error variance σ_U^2 is known

$$\hat{\beta} = \hat{\beta}_{na} \cdot \frac{Var(X^*)}{Var(X^*) - \sigma_U^2}. \tag{16.15}$$

Since the X^* data are available, $Var(X^*)$ is estimated by the corresponding sampling variance. Note that the variance of $\hat{\beta}$ is increased by the inverse reliability compared to $\hat{\beta}_{na}$.

The considerations about the simple model can be extended to the multiple regression model by using the vector notation for β and replacing the variances of X and U by the corresponding variance matrices. With correlated covariates, there is also an effect of measurement error on coefficients of variables, which are measured correctly. This effect is not necessarily an attenuation, but can go in each direction.

16.4 Two approximate methods for handling measurement error

We present two rather simple and popular approximate methods for handling measurement error.

16.4.1 Regression calibration

The first method was introduced by Rosner *et al*. (1989) for logistic regression and established as a general method by Carroll and Stefanski (1990). It works for many regression models with continuous measurement error in the predictors. Due to its easy use, it is one of the reference methods for measurement error correction. The method includes three steps:

1. Find a regression model $E(X|X^*, Z)$ by validation data or replication for the unknown true variable X depending on the observations X^* and further covariates Z.

2. Replace the i*th* observation of the unknown X_i by its predicted values $E(X_i|X_i^*, Z_i)$ and fit the main model.

3. Estimate the variance of the estimator by bootstrap or (complex) asymptotic expansions.

Note that in the first step a model for X given X^* has to be found. However, in many cases a measurement model relates the measured values X^* to X. By making assumptions about the distribution of X and using Bayes' theorem (see Chapter 18), the expected value of X given X^*, $E(X|X^*)$, can be derived.

If a validation study with a gold standard is available, then one can fit a regression model regressing the true values X on X^* and further covariates.

Because the regression calibration method is based on a linear approximation of the regression function, one has to be careful using it in highly nonlinear models.

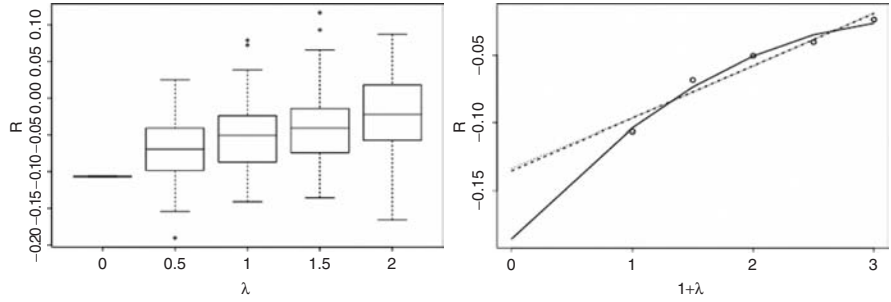


Figure 16.3 SIMEX method. On the left, the simulation step is illustrated. On the right, the extrapolation step with different extrapolation functions is performed.

16.4.2 The simulation and extrapolation approach (SIMEX)

The simulation and extrapolation approach is a very general method for measurement error correction in complex regression models. It has been originally developed by Cook and Stefanski (1994) for models with additive measurement error and extended to the case of misclassification by Küchenhoff *et al*. (2006). An R-software package is available (see Lederer & Küchenhoff, 2006).

The basic idea of the SIMEX is to model the relationship between the amount of measurement error or misclassification and the bias of the estimators of interest when ignoring measurement error. In the simulation step of the SIMEX, new pseudo data sets are produced by adding more measurement error to the variables, which are measured with error. Then, the naive analysis is performed producing estimates, that relate to data with different amounts of measurement error/ misclassification. This is repeated e.g. 200 times for each amount of extra measurement error. The simulation step is illustrated in an example form the Signal-Tandmobiel® study. Here, one research question was, whether there is an East-West gradient for caries experience of children. A logistic regression model with different variables, including the x-coordinate (of the geographical location where the child belonged to) was used. Because there is some misclassification in the caries data, by simulation, the misclassification was enlarged by a simulating new artificial caries data. The enlargement is quantified by a parameter λ. In Figure 16.3 (left panel), one can see the boxplots, which consist of naive estimators for the slope parameter of the x-coordinate with different amounts λ of extra misclassification error in the caries variable. One can see a structure in the dependence of the estimator and the amount of misclassification error given by the parameter λ.

In a next step, a parametric curve (e.g. a quadratic polynomial) is fit to the mean of the estimates modeling the relationship between the amount of measurement error and the estimate of interest. This parametric curve is a function $G(\lambda)$, which is now extrapolated back to the point G(-1). This point relates to the case of no misclassification error. Figure 16.3 (right panel) shows the extrapolation step with different extrapolation functions. The SIMEX method is very general flexible tool

for regression models in the presence of additive measurement error in predictor variables and/or misclassification in predictor variable or a binary response variable.

16.5 Outlook and conclusions

Beside the two presented methods there are many other methods for handling measurement error, see Carroll *et al*. (2006). The most important ones are likelihood and Bayesian analysis. In both cases, one has to write down the likelihoods of the main model and of the measurement model. Then, one has to combine the two models. This can be rather demanding because usually complex integration over the unobserved true variables is necessary. As illustrated in Chapter 18 of this book, MCMC methods are a useful tool to overcome these problems and have been successfully applied in oral health research by Mwalili *et al*. (2005) and Lesaffre *et al*. (2009).

In oral health, there are two fields the methods discussed in this chapter deserve attention by the researchers. One is the misclassification problem in diagnosis, especially in caries research. Another important area is nutrition where many variables can only be measured with substantial error. For an interesting recent discussion of measurement error issues in nutritional epidemiology, see Thiebaut *et al*. (2007). Furthermore, misclassification can also be a relevant issue when data from questionnaires are analyzed.

References

Carroll R & Stefanski L (1990) Approximate quasi-likelihood estimation in models with surrogate predictors. *JASA* **85**, 652–63.

Carroll R, Ruppert D, Stefanski L & Crainiceanu C (2006) *Measurement Error in Nonlinear Models*. Chapman & Hall, New York.

Cook J & Stefanski L (1994) Simulation-extrapolation estimation in parametric measurement error models. *JASA* **89**, 1314–28.

Dion N, Cotart JL & Rabilloud M (2007) Correction of nutrition test errors for more accurate quantification of the link between dental health and malnutrition. *Nutrition* **23**(4), 301–7.

Greenland S (1988) Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine* **7**, 745–57.

Gustafson P (2004) *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall, New York.

Hand DJ (2004) *Measurement Theory and Practice. The World through Quantification*. Arnold, London.

Ismail A (2004) Diagnostic levels in dental public health planning. *Caries Research* **38**, 199–203.

Küchenhoff H, Mwalili S & Lesaffre E (2006) A general method for dealing with misclassification in regression: the Misclassification SIMEX. *Biometrics* **61**, 85–96.

Kuha J, Skinner C & Palmgren J (2001) Misclassification error In *Encyclopedia of Biostatistics* (ed. Armitage P & Colton T), John Wiley & Sons, Ltd, Chichester pp. 2615–21.

Lederer W & Küchenhoff H (2006) A short introduction to the SIMEX and MCSIMEX. *R News* **6**(4), 26–31.

Lesaffre E, Küchenhoff H, Mwalili SM & Declerck D (2009) On the estimation of the misclassification table for finite count data with an application in caries research. *Statistical Modelling* p. accepted.

Morrissey M & Spiegelman D (1999) Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics* **55**, 338–44.

Mwalili S, Lesaffre E & Declerck D (2005) A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. *Applied Statistics* **54**, 77–93.

Neuhaus J (1999) Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* **86**, 843–55.

Pepe M & Janes H (2007) Insights into latent class analysis of diagnostic test performance. *Biostatistics* pp. 474–84.

Rosner B, Willett W & Spiegelmann D (1989) Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* **8**, 1051–70.

Thiebaut A, Freedman LS, Corroll RJ & Kipnis V (2007) Is is necessary to correct for measurement error in nutritional epidemiology?. *Annals of Internal Medicine* **146**, 65–7.

Vanobbergen J, Martens L, Lesaffre E & Declerck D (2000) The Signal-Tandmobiel® project – a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry* **2**, 87–96.

17

Statistical genetics

Amy D. Anderson

17.1 Introduction

One of the main goals of the field of statistical genetics is to discover the genetic variants that contribute to disease. By finding the genetic factors underlying differences in disease risk between individuals, we hope to shed light on the fundamental molecular causes of disease and the biological mechanisms that contribute to disease susceptibility and resistance.

It has been suggested that many oral health diseases have a genetic component. Wilkie and Morriss-Kay [1] provide a review that highlights some of the genetic factors involved in craniofacial development and malformation. A number of these conditions (e.g. Treacher-Collins syndrome [2]) have a simple genetic basis. Even conditions that are often thought of as being sporadic can be influenced by genetics. For example, although hemifacial microsomia seems to be caused by a sporadic event, namely haemorrhaging near the stapedial artery during embryonic development [3], mouse studies have shown that a genetic predisposition for this vessel to rupture can exist [4]. The case for a genetic influence in the development of hemifacial microsomia is also supported by an occasional tendency for the condition to be passed within families [5, 6].

Early-onset periodontitis has been known for some time to have genetic risk factors [7, 8], and, more recently, many studies have investigated associations between various aspects of adult periodontitis (e.g. severity of adult periodontitis, incidence of chronic aggressive periodontitis, incidence of aggressive periodontitis) and immune-response genes [9–17]. Other types of genes have also been suspected of influencing adult periodontitis. For example, a number of genomic

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

sites within the vitamin D receptor gene (VDR) have shown associations with aspects of periodontitis including generalized aggressive periodontitis itself [18], clinical attachment loss due to periodontal disease [19] and periodontal disease progression [20]. A recent pilot study found a preliminary association between genes relating to the body's metabolism of xenobiotics, and chronic and aggressive forms of periodontal disease [21].

Studies have also been done investigating the genetic risk factors for oral cancer. Several genes involved in the metabolism of carcinogens have shown associations with the development of oral cancer or oral precancerous lesions [22–24]. Kuroda *et al*. [25] found associations between genetic variants at a tumor-suppressor gene and both (a) risk of oral cancer and (b) post-treatment prognosis.

The above-mentioned studies used a variety of methods to look for genetic associations. The common methods for using genetic data to either find genes contributing to an outcome (i.e. disease risk) or test whether a gene might influence a trait or outcome can be divided into two main categories: Family-based studies that look to see whether specific genetic variants tend to be passed from parent to offspring along with disease status (or, alternatively, tend to be passsed independently of disease status), and population-based studies that look for associations between individual's genetic make-up at specific chromosomal locations and the individual's disease status or trait value.

In this chapter, we begin with a review of basic genetics, explore a few concepts in genetics that are of interest in gene mapping studies, then give an introduction to issues surrounding the use of a particular population-based design: the case-control genetic association study.

17.1.1 Genetic data

A person's genetic information is encoded in deoxyribonucleic acid (DNA). For our purposes, we can think of a piece of DNA as a string of nucleotides. There are four possible types of nucleotides: adenine (A), guanine (G), cytosine (C), and thyamine (T). Chromosomes are packaged strings of DNA. There are two main types of chromosomes possessed by humans: *autosomes* are chromosomes that follow the usual rules of inheritance – an individual has two copies of each autosome, one inherited from each of the individual's two parents; and *sex chromosomes* include the X and Y chromosomes that determine an individual's sex. The human genome, the collection of all hereditary information possessed by an individual, would then include the 22 autosomes and two sex chromosomes, as well as a small amount of mitochondrial DNA. Although risk of disease can be effected by DNA on any of these chromosome types, most of an individual's genetic material is contained in the autosomes, so we will focus exclusively on the autosomes for the remainder of this chapter.

The word *locus* (plural *loci*) is used to refer to a small segment of DNA on a chromosome. Often, different copies of the same chromosome vary in their content at a locus. A locus that exhibits such variation is called a *polymorphism*. The variants themselves are called *alleles*. Because each individual has two copies

of each chromosome (remember, we are focusing on the autosomes), he or she has two alleles at any locus. This pair of alleles constitute the individual's *genotype* for that locus. If the two alleles are of like type, the individual is said to be *homozygous*, otherwise, he or she is said to be *heterozygous*. For example, a particular locus might have alleles *A* and *a*, so an individual might have genotype *AA*, *Aa*, or *aa*, depending upon which allele is on each of the individual's two chromosomes. Continuing with this example, *AA* and *aa* individuals are homozygous whereas *Aa* individuals are heterozygous.

A genetic *marker* is a locus at which we can observe an individual's genotype. The most common type of genetic marker used in case-control studies is a *single nucleotide polymorphism* (SNP – pronounced 'snip'). A SNP is a site on a chromosome at which individuals vary at a single nucleotide. Although, in theory, it is possible for four different alleles to exist in a population (namely, the four possible nucleotides), it is actually quite rare for more than two of the four possible alleles to occur at any given site in any population (or even among all of mankind). In other words, SNPs are nearly always *diallelic* (i.e. have two possible alleles). SNPs are very common and there will usually be multiple SNPs within a gene. This abundance makes SNPs the marker of choice for researchers hoping to find specific variants within a gene that contribute to disease risk.

17.1.2 Mendel's laws, recombination, and linkage

Mendel's first law states that, at a single locus, a parent randomly chooses one of its two alleles to pass to an offspring, independent of which allele it has passed to any other offspring. For example, if an individual with genotype *Aa* has a spouse with genotype *Aa*, then any child from this union will have genotypes *AA*, *Aa*, or *aa* with probabilities 1/4, 1/2, and 1/4, respectively (since each parent is equally likely to pass an *A* or an *a* allele).

Mendel's second law deals with inheritance at two or more loci. This law states that which allele a parent passes at one locus is independent of which of that parent's alleles is passed at any other locus. This law is true for loci that lie on different chromosomes, but is not true in general. In fact, alleles at loci that lie close together on a chromosome are likely to be passed together. Figure 17.1 gives a schematic view of the process by which an individual creates, from his/her two copies of a chromosome, a single chromosome to be passed to the next generation. The key feature is that, during meiosis, *crossovers* can occur at sites where lined-up maternal and paternal chromosomes touch (points 1, 2, 3, and 4 in the figure), and, when this happens, the chromosomes exchange material. As a result, a chromosome to be passed to the next generation will consist of alternating segments of maternal and paternal type.

As an example, consider loci $\mathcal{A}$ and $\mathcal{B}$, where locus $\mathcal{A}$ has alleles *A* and *a* and locus $\mathcal{B}$ has alleles *B* and *b*. Suppose an individual inherited alleles *A* and *B* from his mother and alleles *a* and *b* from his father. If the loci are on different chromosomes, the individual will pass alleles *AB*, *Ab*, *aB*, and *ab*, each with probability 1/4. Alternatively, if the loci lie so close together on a chromosome

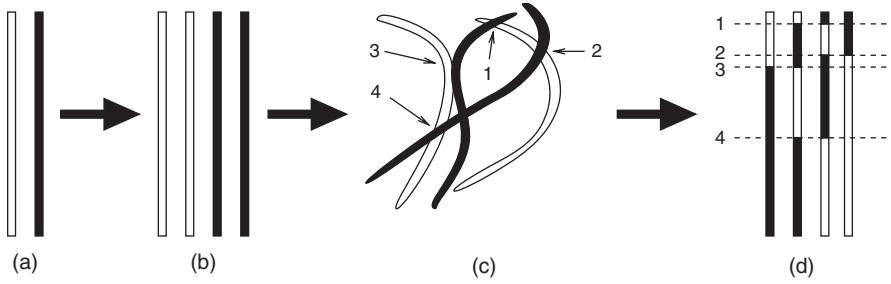


Figure 17.1 Crossovers during meiosis. (a) An individual has two copies of each chromosome (one inherited from each parent). (b) During meiosis, a second copy is made of both the maternal and paternal chromosomes. (c) Chromosomes lie loosely in the cell and can exchange material at places where they touch (here, points 1, 2, 3, and 4). (d) After exchange, there are four possible chromosomes that could be passed to the following generation. One of these will be chosen at random to be passed to the next generation.

that there is essentially no chance that a crossover could occur between them, then the individual will pass alleles $\{A, B\}$ and $\{a, b\}$, each with probability 1/2. In the general case, the individual will pass alleles AB, Ab, aB, and ab, with probabilities $\frac{1-\theta}{2}$, $\frac{\theta}{2}$, $\frac{\theta}{2}$, and $\frac{1-\theta}{2}$, respectively, where θ is the probability that a *recombination* has occured between the loci, that is, that the genetic material passed at locus A has a different parental source than the genetic material passed at locus B. Since each crossover switches the parental source of the genetic material, an odd number of crossovers between two loci results in a recombination whereas, if an even number of crossovers occurs, both loci will have the same parental source. For loci on different chromosomes or very far apart on the same chromsome, $\theta = 0.5$. For loci on the same chromosome, θ is the probability that an odd number of crossovers occurs on the chromosome between loci $\mathcal{A}$ and $\mathcal{B}$ in the meiosis of interest. In either case, θ is called the recombination frequency between the loci.

Two loci are said to be *linked* if $\theta < 0.5$ and *unlinked* if $\theta = 0.5$. Note that this has to do with dependence between the parental origin of the genetic material passed at the two sites *in a single meiosis*. This is often confused with *linkage disequilibrium*, which has to do with dependence between the alleles present at two sites in a random chromosome drawn from the population, which we will discuss in the next section.

17.2 Dependence between alleles

Statistical dependence between alleles is of interest for a number of reasons, not the least of which is that assumptions regarding this dependence lie at the heart of many gene-mapping methods. The two main types of dependence that will be of interest are Hardy-Weinberg equilibrium/disequilibrium, which refers to independence/

dependence between an individual's two alleles at a single locus, and linkage equilibrium/disequilibrium, which refers to independence/dependence between alleles at two different loci. Both types of dependence are described below.

17.2.1 Hardy-Weinberg equilibrium

Consider a single locus, $\mathcal{A}$, that has alleles A and a with frequencies p_A and p_a, respectively. If an individual's two alleles at $\mathcal{A}$ are two randomly chosen alleles from the population, then the genotype frequencies are:

$$\begin{aligned}\Pr(AA) &= p_A^2 \\ \Pr(Aa) &= 2p_A(1-p_A) \\ \Pr(aa) &= (1-p_A)^2.\end{aligned}$$

The two in the second equation comes from the fact that an individual could obtain genotype A_iA_j in two ways: either the individual received allele A_i from his mother and A_j from his father or conversely.

Hardy-Weinberg equilibrium (HWE) will hold at a locus in any population in which mating is done at random with respect to that locus. Deviations from Hardy-Weinberg equilibrium can be caused by non-random mating. As an example, suppose that a population consists of two subpopulations of equal size in which individuals always choose mates from within their subpopulation, but within each subpopulation mates are chosen at random. Consider a locus with two alleles, A and a, where the frequency of allele A is 0.8 in subpopulation 1 and 0.2 in subpopulation 2. Table 17.1 shows the genotype frequencies in this situation. Note that, in the overall population, the frequency of allele A is 0.5, so, if the population was in Hardy-Weinberg equilibrium, the genotype frequencies would be 0.25, 0.5, and 0.25 for genotypes AA, Aa, and aa, respectively. The actual proportions are 0.34, 0.32, and 0.34. In this case, population structure (the fact that individuals are not mating at random but, rather, chosing mates from within their own subpopulation) has resulted in a deviation from HWE.

As a general rule, if allele frequencies vary between different subpopulations and individuals tend to choose mates within their own subpopulation, the result

Table 17.1 Genotype frequencies in a population consisting of two equally sized subpopulations where mating is done at random within a subpopulation but mating does not occur between subpopulations. In this example, allele A has frequency 0.8 and 0.2 in subpopulations 1 and 2, respectively.

Genotype	Subpopulation 1	Subpopulation 2	Total population
AA	0.64	0.04	0.34
Aa	0.32	0.32	0.32
aa	0.04	0.64	0.34

will be a deviation from Hardy-Weinberg. The dependence between an individual's alleles can be thought of this way: Once it is observed that an individual's paternal allele is, say, an *A* allele, then there is an increased probability that the individual's father might be from a subpopulation where the *A* allele is relatively common. Then, since the individual's mother is likely to have come from the father's subpopulation, she has an increased probability of also passing the *A* allele (relative to a mother randomly chosen from the population at large).

Genotyping errors can also cause the appearance of deviations from Hardy-Weinberg equilibrium. If there are serious problems with genotyping at a particular marker, this might show up in the data as an excess of homozygous, or, more rarely, heterozygous individuals, and, hence, as a deviation from Hardy-Weinberg proportions.

17.2.1.1 χ^2 test for Hardy-Weinberg equilibrium

Testing for Hardy-Weinberg equilibrium is generally done via a χ^2 test or by Fisher's exact test (see Chapter 10 for a discussion of these tests in a general context). The null hypothesis being tested is that the population is in HWE. We now outline the procedure for χ^2 testing with a diallelic locus. The procedure for multiple alleles is similar. Suppose we have a random sample of n individuals from the population. This sample then contains $2n$ alleles. Let n_{AA}, n_{Aa} and n_{aa} denote the number of individuals in our sample with genotypes *AA*, *Aa*, and *aa*, respectively. From these, we can obtain maximum likelihood estimates for the allele frequencies as follows:

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n}$$

$$\hat{p}_a = \frac{2n_{aa} + n_{Aa}}{2n}.$$

Given these allele frequencies, we can derive the expected genotypic counts under HWE:

$$E[n_{AA}] = n\hat{p}_A^2$$

$$E[n_{Aa}] = 2n\hat{p}_A\hat{p}_a$$

$$E[n_{aa}] = n\hat{p}_a^2.$$

We can test for Hardy-Weinberg equilibrium using a test statistic with the form

$$X^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

$$= \frac{[n_{AA} - n\hat{p}_A^2]^2}{n\hat{p}_A^2} + \frac{[n_{Aa} - 2n\hat{p}_A(1-\hat{p}_A)]^2}{2n\hat{p}_A(1-\hat{p}_A)} + \frac{[n_{aa} - n(1-\hat{p}_A)^2]^2}{n(1-\hat{p}_A)^2}.$$

This test statistic follows an approximate χ^2 distribution with 1 degree of freedom, provided that the sample size is sufficiently large. The null hypothesis (that the

population is in Hardy-Weinberg equilibrium) is rejected if the value of the test statistic exceeds a cutoff determined by the required size of the test. For example, if we are to have a Type 1 error rate of $\alpha = 0.05$, we find the 95th percentile of the χ^2_1 distribution, namely, 3.84. We reject the hypothesis of Hardy-Weinberg equilibrium if our test statistic exceeds the value of 3.84.

As an example, a data set consisting of 125 men that were collected as part of Inagake *et al.*'s study [20] on the relationship between genotypes at a site within the vitamin D receptor gene and periodontal disease progression had the following genotypes: 41 *AA*, 58 *Aa*, and 26 *aa*. The estimated allele frequencies are $\hat{p}_A = 0.56$ and $\hat{p}_a = 0.44$. The expected genotype counts are then $E[n_{AA}] = 39.2$, $E[n_{Aa}] = 61.6$, and $E[n_{aa}] = 24.2$. The test statistic is $X^2 = 1.896$. This dataset does not suggest a departure from Hardy-Weinberg Equilibrium. The P-value for this example is $p = \Pr(X^2 \geq 1.896 | X^2 \sim \chi^2_1) = 0.169$. Since this P-value is fairly large, we conclude that this marker is neigher showing obvious signs of systematic genotyping errors that might cause a departure from HWE nor exhibiting evidence of population structure. More information about interpreting the results of Hardy-Weinberg testing are given in Section 17.2.1.3.

As a rule of thumb, we might use this test if the expected count in each cell is above 5 or so. If this fails (as might happen if one of the alleles is fairly rare and your sample is small) the χ^2 test is inappropriate and significance is obtained by Fisher's exact test.

17.2.1.2 Fisher's exact test for Hardy-Weinberg equilibrium

Another way to attach a P-value to our sample, that does not depend upon our having a sufficiently large sample to justify using the χ^2 distribution, is to return to the definition of P-value. The P-value associated with a sample is the probability that you would have obtained a sample *at least as contrary to the null hypothesis* as the sample you obtained, if the null hypothesis is true. To calculate this value, we sum the probabilities of all samples that are 'at least as contrary to the null hypothesis' as the sample we obtained. In other words, we sum the probabilities of all samples that are at least as improbable as our own sample. Since we do not want any part of our test to be a test of what the allele frequencies are in the population, we perform all calculations conditioning on the number of alleles of each type seen in the data. If Hardy-Weinberg equilibrium holds, the probability of a data set with counts n_{AA}, n_{Aa}, and n_{aa} is

$$\begin{aligned}\Pr(n_{AA}, n_{Aa}, n_{aa}) &= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(p_A^2)^{n_{AA}}(2p_A p_a)^{n_{Aa}}(p_a^2)^{n_{aa}} \\ &= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}2^{n_{Aa}}p_A^{n_A}p_a^{n_a},\end{aligned}$$

where n_A and n_a represent the numbers of A and a alleles observed in our sample, respectively. We will want to condition on the observed allele counts, so we note that the probability of observing our allele counts is $\Pr(n_A, n_a) = \frac{(2n)!}{n_A!n_a!}p_A^{n_A}p_a^{n_a}$.

Table 17.2 Possible genotype counts (n_{AA}, n_{Aa}, n_{aa}) and probabilities for samples with 10 *A* and 24 *a* alleles. The bold-face sample represents the genotype counts for 17 patients with severe chronic periodontitis in a study by Gonzales *et al*. [26]. The P-value is the sum of all probabilities in this table less than or equal to 0.152. Hence, the P-value is 0.241.

Genotype counts	Probability	Genotype counts	Probability
(5, 0, 12)	4.7×10^{-5}	(2, 6, 9)	0.332
(4, 2, 11)	5.7×10^{-3}	(1, 8, 8)	0.427
(3, 4, 10)	0.083	**(0, 10, 7)**	**0.152**

Then

$$\Pr(n_{AA}, n_{Aa}, n_{aa}|n_A, n_a) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}2^{n_{Aa}}p_A^{n_A}p_a^{n_a}\cdot\frac{n_A!n_a!}{(2n)!}p_A^{-n_A}p_a^{-n_a}$$
$$= \frac{n!}{(2n)!}\cdot\frac{n_A!n_a!}{n_{AA}!n_{Aa}!n_{aa}!}2^{n_{Aa}}.$$

Using the above formula, we can calculate the probability of each possible data set and sum the probabilities of all such datasets that are less likely than the actual dataset we obtained.

As an example, a study conducted by Gonzales *et al*. [26] that looked at the relationship between periodontitis and genetic variation in the IL-10 gene included 17 individuals with severe chronic periodontitis. At a particular locus, this sample contained no *AA* individuals, 10 *Aa* individuals, and 7 *aa* individuals. In this case, there are a total of 10 *A* alleles and 24 *a* alleles in the sample. Table 17.2 gives a list of genotype counts and probabilities for all possible datasets with these allele counts. In this example, the P-value is 0.241.

17.2.1.3 Interpreting the results of tests for Hardy-Weinberg equilibrium

Most gene-mapping methods rely on an assumption of Hardy-Weinberg equilibrium. Violations of this assumption may be the result of a number of factors including ascertainment issues (e.g. at a locus that has a strong influence on disease risk, a set of case individuals will be enriched for certain genotypes), nonrandom mating (e.g. the presence of population structure), or problems with the genotyping procedure at a locus. In any of these cases, a deviation from Hardy-Weinberg equilibrium is of interest.

Many researchers begin their data analysis by testing for Hardy-Weinberg equilibrium at each of their markers. If the data can be treated as a random sample of individuals (as would be the case in a prospective study), interpretation of the test is straight-forward: Deviations from HWE at many markers may be indicative of population structure (which can be a problem if you intend to do a population-based case-control study), whereas big deviations at a single marker can indicate problems with the genotyping at that marker.

With case-control data, it is difficult to know how to perform the tests for HWE. One should not perform a test on the combined case-control sample as this sample will not look like a random sample from the population (it will almost certainly be enriched for cases relative to the general population). In this case, testing for Hardy-Weinberg equilibrium can be done in the cases and controls separately, but the researcher must interpret the results in the context that deviations from HWE at a marker can be caused by association with the disease. In this case, deviations from HWE at more markers than could reasonably be expected to be associated with the disease might be an indication of population structure. Deviations at a single marker are difficult to interpret directly: the deviation could be caused by the type of association the study is designed to find, or the deviation could be indicative of genotyping errors.

17.2.2 Linkage disequilibrium

Linkage disequilibrium (LD) refers to a lack of independence between alleles at different loci. Association studies aimed at gene mapping look for associations between an individual's case-control status (or trait value, if we are looking for a gene that influences a trait, such as severity of periodontitis, that we might want to treat as a quantitative variable) and his/her alleles at a genetic marker. The ideal situation is when the locus that contributes to disease susceptibility is among those that have genotyped. It is more often the case, however, that the risk locus itself isn't genotyped, but a nearby marker is. If alleles at the genotyped marker are in LD with alleles at the risk locus, then associations between case-control status (or trait values) and (unobserved) alleles at the risk locus will translate into associations between marker case-control status (or trait values) and observed alleles at the marker.

17.2.2.1 Measures of linkage disequilibrium

Consider loci $\mathcal{A}$ and $\mathcal{B}$ with alleles A and a, and B and b, respectively. Let p_A be the frequency of allele A in the population, and define p_a, p_B, and p_b similarly. Let p_{AB} be the probability that a random chromosome from the population has alleles A and B at loci $\mathcal{A}$ and $\mathcal{B}$. Define p_{Ab}, p_{aB}, and p_{ab} similarly. If there is linkage equilibrium (i.e. independence) between the two loci, then $p_{AB} = p_A p_B$. The most basic measure of LD between alleles A and B is $D_{AB} = p_{AB} - p_A p_B$ (the difference between the proportion of chromosomes that have alleles A and B and what that proportion would be under independence). D_{AB} is also often seen written as $D_{AB} = p_{AB}P_{ab} - p_{Ab}p_{aB}$. That the two forms are equivalent is shown in the following derivation:

$$\begin{aligned} D_{AB} &= p_{AB} - p_A p_B = p_{AB} - (p_{AB} + p_{Ab})(p_{AB}p_{aB}) \\ &= p_{AB} - p_{AB}(p_{AB} + p_{aB} + p_{Ab}) - p_{Ab}p_{aB} = p_{AB} - p_{AB}(1 - p_{ab}) - p_{Ab}p_{aB} \\ &= p_{AB}p_{ab} - p_{Ab}p_{aB}. \end{aligned}$$

In this second form, it is easy to see that $D_{AB} = D_{ab} = -D_{Ab} = -D_{aB}$. Also, as a technical note, if X is an indicator variable for allele A and Y is an indicator for allele B, then D is the covariance between X and Y (which is the correlation multiplied with the standard deviation of see X and Y, see also Chapter 11).

The difficulty with D as a measure of association is that it is difficult to interpret. For loci $\mathcal{A}$ and $\mathcal{B}$ above, suppose that $p_A = 0.01$ and $p_B = 0.5$. If allele A only appears on chromosomes with allele B, we have $\Pr(B|A) = 1$, so the association is in some sense as strong as possible. In this case, $p_{AB} = p_A = 0.01$, so $D = p_{AB} - p_A p_B = 0.01 - (0.01)(0.5) = 0.005$. Next, consider the case where $p_A = 0.5$ and $p_B = 0.5$. Suppose there is a weak association between alleles A and B, where, say, $\Pr(B|A) = 0.51$ (just slightly larger that $\Pr(B) = 0.5$, so A is slightly more likely to appear on a chromosome with B than would be the case under independence). Then $p_{AB} = 0.255$, and $D = 0.255 - (0.5)(0.5) = 0.005$. Hence, $D = 0.005$ could indicate either a strong or a weak association.

One solution, proposed by Lewontin [27], is to scale D by the maximum possible value D could obtain for alleles with the given frequencies. The maximum possible value of D is

$$D^* = \begin{cases} \min\{p_A p_b, p_a p_B\} & \text{if } D \geq 0 \\ \min\{p_A p_B, p_a p_b\} & \text{if } D < 0. \end{cases}$$

Lewontin's measure of LD is $D' = \frac{D}{D^*}$. This measure of LD will be 1 or -1 in the case where only three of the four possible combinations of alleles (AB, Ab, aB, and ab) exist within the population. In other words, $|D| = 1$ when one allele always appears with a particular allele at the other locus (e.g. if A always appears on a chromosome with B).

A second solution is to simply use the correlation between the indicator variables for alleles A and B. This measure, proposed by Hill and Weir [28] is $r = \frac{D}{p_A p_a p_B p_b}$. This measure has the property that $|r| = 1$ when only two possible combinations of alleles (AB, Ab, aB, and ab) exist in the population, that is, if knowledge of the allele present at either locus completely determines the allele at the other locus.

17.2.2.2 Causes of linkage disequilibrium

In order for a genetic association study to work, genotypes must be observed either for a 'causal' locus that influences the trait/disease of interest, or at a nearby marker that is in linkage disequilibrium with the causal locus. In this section, we outline an argument for why we believe loci that are in close proximity to each other should be in linkage disequilibrium with each other (a situation we hope to exploit in gene-mapping studies), then look at an alternate scenario in which LD can be generated between loci that are not necessarily located close to each other (a situation that can be disastrous to association studies).

Let us consider a simple situation in which there are two loci, $\mathcal{A}$ and $\mathcal{B}$ located on a chromosome. Suppose that allele A at $\mathcal{A}$ confers some risk of disease. We

can imagine a time before the A allele ever existed, when all chromosomes had allele a at the $\mathcal{A}$ locus and had either allele B or b at the $\mathcal{B}$ locus. At this time, all chromosomes had the form aB or ab. At some point in time, a mutation creates allele A. The chromosome with this new mutation had either a B or a b allele. Without loss of generality, suppose that chromosome had a B allele. At this point in time we have complete LD between alleles A and B (in the sense that there are three possible versions of the chromosome and $D'_{AB} = 1$). Now imagine how allele A might be passed to the next generation: if no recombination occurs between loci $\mathcal{A}$ and $\mathcal{B}$ in the meiosis, the chromosome that passes A to the next generation will have the B allele at locus $\mathcal{B}$. If a recombination does occur, the chromosome being passed will be of type AB if the individual's other copy of the chromosome also had the B allele (which will happen with probability p_B), or will be of type ab if the other copy of the chromosome had the b allele.

Next, we will look at how linkage disequilibrium changes through the generations. Our model will be that, in any generation, each individual's two alleles at a locus are two random alleles chosen (with replacement) from the previous generation. In this scheme, if the population is sufficiently large, allele frequencies will change very little from generation to generation and can be treated as being constant through time. Let $D(t)$ be the amount of linkage disequilibrium (measured as D described above) between loci $\mathcal{A}$ and $\mathcal{B}$ in generation t and let θ be the recombination frequency between the two loci. If $p_{AB}(t)$ and $p_{Ab}(t)$ represent the frequencies of chromosomes AB and Ab in generation t, respectively, then we can find $p_{AB}(t+1)$ as follows:

We want to know whether a random copy of this chromosome drawn in generation t has alleles A and B. We find this by conditioning on whether or not the meiosis that produced that chromosome experienced a recombination between the two loci. If it did not, then the chromosome is of type AB if it was descended from a chromosome of type AB. If a recombination did occur, then the chromosome was descended from a chromosome with allele A at the $\mathcal{A}$ locus and another chromosome that had allele B at the $\mathcal{B}$ locus. Then

$$p_{AB}(t+1) = (1-\theta)p_{AB}(t) + \theta p_A p_B.$$

subtracting $p_A p_B$ from both sides yields $D_{AB}(t+1) = (1-\theta)D_{AB}(t+1)$, so

$$D_{AB}(t) = (1-\theta)^t D_{AB}(0).$$

This derivation can be found in Lynch and Walsh [29]. The result is that, under random mating, linkage disequilibrium decreases by a factor of $(1-\theta)$ per generation. Return now to the example in which mutation created a new risk allele that was in strong LD ($D' = 1$) with all other alleles on its chromosome. If we imagine that the risk allele is now very old, we see that recombination would have decreased that initial LD to imperceptible levels for all alleles except those belonging to loci which lie so close to the site of the mutation that $\theta \approx 0$.

Note that there was nothing in the above argument that required the new allele to be a risk allele – any new allele will begin its existence in strong LD with

alleles at nearby loci. Over the generations, recombinations will gradually destroy LD so that, in a randomly mating population, only alleles that are located very close together show LD. This agrees with what is generally observed: Alleles at loci that are far apart look like independent alleles from the population, but alleles at loci that are very close together show strong dependence.

There is often confusion between linkage and linkage disequilibrium. Both have to do with dependence between alleles at different loci. Linkage has to do with independence between alleles passed from parent to child *in a single meiosis*. The scale at which linkage exists is large: two loci may have to be about 70 million nucleotides (70 Mb) apart to be roughly unlinked ($\theta < 0.4$). Linkage disequilibrium has to do with independence between alleles at two loci in a random chromosome drawn from the population. Hence, it gets to the question, 'Have there been enough recombinations between these two loci, *in all the generations since these alleles were created*, to break down the LD that existed between them when these alleles first came into existence?' Linkage disequilibrium is usually a fine-scale phenomenon. Weir *et al*. [30] showed a plot in which they looked at the decay of LD (measured by r^2) as a function of inter-marker spacing for markers in the HapMap data set [31]. They found that, for most populations examined, $r^2 < 0.1$ when the markers examined were separated by about 50 intervening markers. Since these markers were at an average spacing of roughly 0.005 Mb, we see that LD is low ($r^2 < 0.1$) for loci that are about 0.27 Mb apart.

Tight linkage between loci slows the decay of LD between them, so it is tempting to assume that loci must be linked to be in LD. This is not true. Associations between alleles at different loci can be caused by population structure. Suppose that a population consists of two equally sized subpopulations, and alleles A and B are common in subpopulation 1 (say, $p_A = p_B = 0.9$) but rare in subpopulation 2 (say, $p_A = 0.2$, $p_B = 0.1$). You may imagine that these loci are completely unlinked – they sit on different chromosomes, and are not in LD in either subpopulation. A random chromosome drawn from the first population will be of type AB with probability 0.81. A random chromosome from the second population will be of type AB with probability 0.02. In the population at large, then, $p_{AB} = 0.415$, and, again in the population at large, $p_A = 0.55$ and $p_B = 0.5$. Then, $D' = [0.415 - (0.55)(0.5)]/0.225 = 0.622$. In this situation, we have completely unlinked loci that exhibit LD.

17.3 Genetic case-control studies

There are many study designs and methods of analysis that can be used to detect linkage and/or association between a marker and a locus that contributes to disease risk (or quantitative trait value). To give a flavor of one possible design, we will consider the case-control association study (a discussion of case-control studies in a more general context appears in Chapter 7). Once the decision has been made about the type of design to be used in selecting the human subjects, the researcher has to consider a second question: which markers to include in the

study. The simplest situation is when the researcher has biological reasons for limiting the study to genetic markers within a moderate number of 'candidate' genes. An alternate situation is for genotypes to be recorded at a dense set of markers located throughout the genome. At the current time, studies are being performed using panels of 500K SNP markers (e.g. Wellcome Trust Case Control Consortium [32]) and Affymetrix, a popular genotyping company, is advertising a panel of 1.8 million SNPs. Both the 'candidate gene' and 'dense panel' approaches have pros and cons: The candidate gene approach will miss any association that is not in a gene previously suspected to be involved with the disease. The dense panel approach suffers from difficulties with multiple testing – only loci with very strong effects will be able to survive a multiple testing correction unless the sample size is extraordinarily large.

17.3.1 Tests for association

Testing at an individual marker is the same regardless of how that marker was chosen. The most common tests are all χ^2 tests. Suppose we are looking at a marker with alleles A and a. Define n^C_{AA}, n^C_{Aa}, and n^C_{aa} to be the number of case individuals with genotypes AA, Aa, and aa, respectively and define n^N_{AA}, n^N_{Aa}, and n^N_{aa} to be the corresponding value for the control (normal) individuals. Let n_C and n_N be the number of case and control individuals, respectively, and let $T = n_C + n_N$ be the total number of individuals in the study.

The *allelic test* tests for whether allele counts are different in cases and controls. This is the usual χ^2 test for independence based on the contingency table shown in Table 17.3. The test has one degree of freedom. By looking at allele counts instead of genotype counts, we are making an assumption that our data consists of random alleles from the population. There is an inherent assumption of Hardy-Weinberg equilibrium in this test, as shown by Sasieni (1997).

An alternative is the *genotypic test* in which the test is performed on the genotypes. This two degree of freedom test is the usual χ^2 test of independence based on the contingency table given in Table 17.4. It is generally prefered to the allelic test because it does not require an assumption of Hardy-Weinberg equilibrium.

The genotypic test examines whether genotype frequencies differ between cases and controls, but makes no assumptions as to the form of that difference. If one is willing to assume a model for how genotypes relate to case-control status, a more powerful test can be derived. If we assume a linear relationship between the probability of being a case and the number of a alleles carried by an individual,

Table 17.3 Contingency table for the allelic association test.

	A	a
Case	$2n^C_{AA} + n^C_{Aa}$	$2n^C_{aa} + n^C_{Aa}$
Control	$2n^N_{AA} + n^N_{Aa}$	$2n^N_{aa} + n^N_{Aa}$

Table 17.4 Contingency table for the genotypic association test.

	AA	Aa	aa	Total
Case	n_{AA}^C	n_{Aa}^C	n_{aa}^C	n_C
Control	n_{AA}^N	n_{Aa}^N	n_{aa}^N	n_N
Total	n_{AA}	n_{Aa}	n_{aa}	T

then the Cochran-Armitage test statistic Agresti:[34] for this situation is

$$X_A^2 = \frac{T\left[T(n_{Aa}^C + 2n_{aa}^C) - n_C(n_{Aa} + 2n_{aa})\right]^2}{n_C n_N\left[T(n_{Aa} + 4n_{aa}) - (n_{Aa} + 2n_{aa})^2\right]}.$$

This statistic has an asymptotic χ^2 distribution with 1 degree of freedom.

As an example, we will consider a portion of the data presented in Nibali *et al.*'s study of aggressive periodontitis and SNPs in the promoter region of the Interleukin-6 gene [15]. Table 17.5 shows the genotypes and allele counts at a SNP 1480 basepairs before the IL-6 gene for 220 case and 229 control individuals with European ancestry. In this case the allelic test yields a χ^2 value of 12.32, which gives a P-value of 4.5E-4. The genotypic test gives a test statistic of 11.71 and a P-value of 0.003. The Cochran-Armitage test statistic is

$$X_A^2 = \frac{449\left[449(59 + 2\cdot 14) - 220(146 + 2\cdot 39)\right]^2}{220\cdot 229\left[449(146 + 4\cdot 39) - (146 + 2\cdot 39)^2\right]} = 10.89.$$

The P-value is 9.7E-4. This association is significant for all three tests.

More generally, these analyses can be framed in terms of a logistic regression of case status on the number of a alleles possessed by an individual. That framework can be used to incorporate covariates (e.g. age, sex, etc.) into the model.

When interpreting the results of these tests, the researcher should keep in mind that systematic genotyping errors can cause false associations between markers and disease status. The particular case to keep in mind is when, as is unfortunately often the case, the case individuals have their genotypes evaluated separately (at

Table 17.5 Contingency tables for the genotypic association test using Nibali's data [15].

	Genotypic Table					Allelic Table		
	CC	GC	GG	Total			G	C
Case	147	59	14	220		Case	353	87
Control	117	87	25	229		Control	321	137
Total	264	146	39	449				

a different time or location) than the case individuals. To screen for this type of false positive, it is good practice to confirm positive results by looking at the raw genotyping results that can be obtained from the lab that evaluated the genotypes. For more information on this essential procedure, which involves merely eyeballing the raw data to confirm that the genotypes look plausible, see, for example, the supplementary material for Wellcome Trust Case Control Consortium [32].

17.3.2 Sample size

The power of these tests of association will depend on a number of factors: the strength of the effect at the risk locus, the degree of LD between the risk locus and marker locus, allele frequencies at both risk and marker loci, sample size, and the criterion for rejecting the null hypothesis. The Wellcome Trust Case-Control Consortium performed a simulation study in which they sequentially chose each SNP in a region to be the 'risk' locus, simulated case-control data (2000 cases, 3000 controls) based on that locus, and determined whether the association would be detected at any of the subset of SNPs in the region that happened to be on the Affymetrix 500K SNP panel. This was done for 10 different regions and three different relative risks. The results were described in Supplementary Table 17.2 of the article [32]. They found that, at a relative risk of 1.3, the Cochran-Armitage trend test had a power of 0.46 when the P-value criterion was 1×10^{-6}, corresponding to a type-I error rate of 0.05 with a Bonferroni adjustment for 50000 markers. This increased to 0.81 for a relative risk of 1.5 and 0.91 for a relative risk of 1.7. We include this information to give a feel for the sample size needed to detect a rather strong effect when multiple testing considerations require a stringent P-value to be required.

17.3.3 Population structure

If the population being sampled consists of multiple groups, differences in allele frequencies between the groups can result in associations between unlinked loci. In addition, if the subgroups within the population differ in risk of disease, spurious associations will be expected to arise between case-control status and alleles at any locus at which the allele frequencies vary between groups.

To illustrate this point, Campbell *et al*. [35] showed that a spurious association can be detected between alleles at the LCT gene (which codes for lactase tolerance and has allele frequencies that vary within Europe along a north–south gradient) and height (which also varies in Europe along a north–south gradient) within a sample of Americans of European ancestry. In another study, Knowler *et al*. [36] found a spurious association between certain genetic markers and type 2 diabetes mellitus in Pima Indians. The flaw in this study was that it did not take into account the fact that many of the individuals included in the study had mixed European and American Indian ancestry. The markers in question had different allele frequencies among Europeans and Pima Indians and the risk of diabetes also differed between the two groups. When Knowler *et al*. repeated the study using only Pima Indians without European ancestry, they failed to find any association.

Methods exist to discover and/or correct for hidden population structure within a dataset, but these methods tend to be overly conservative. One popular method, genomic control [37], is based on the premise that, for the Cochran-Armitage trend test, population structure will result in a test statistic that has a null distribution that looks like that of a constant multiple of a χ^2 random variable with one degree of freedom. That is, $Y^2 = \lambda X^2$ for some λ, where Y^2 is the Cochran-Armitage test statistic calculated on a dataset that comes from a population with some structure, and X^2 has a χ^2 distribution with 1 degree of freedom. Under the hypothesis that the vast majority of markers will come from the null distribution, one can examine the distribution of Y^2 values for a large set of markers and determine λ for the dataset. Testing is then done on the statistic Y^2/λ. Studies examining the behavior of Genomic Control have found mixed results. Shmulewitza *et al*. [38] performed a simulation study and found that the method was generally (but not always) conservative.

17.4 Discussion

In this chapter we have introduced some of the basic ideas in statistical genetics including a review of genetic terminology and some coverage of Hardy-Weinberg equilibrium and linkage disequilibrium. We have included a discussion of some methods for testing for genetic association using case-control data and have highlighted some of the challenges (e.g. population structure, genotyping errors) that confront researchers performing this type of study.

Genetic factors play a large role in several oral health diseases, including some (e.g. periodontal disease) with large public health impacts. A number of previous studies have attempted to locate genetic variants associated with oral health disease, but these have, to date, met with limited success. This mirrors the situation with genetic association studies in other fields: until recently, the record for producing reproducible findings was abysmal. Recent advances, including a new awareness of the role of genotyping error in producing false positives (and, subsequently, a flurry of research into improving genotype-calling algorithms), as well as a more realistic view of what constitues an appropriate sample size (mere dozens of cases and controls will not suffice for a genome-wide study), have resulted in a series of successes. In the year 2007 alone, genetic associations with heart disease [39], Crohn disease [40], type 1 diabetes [41], and type 2 diabetes [42], as well as other diseases have been replicated in distinct datasets.

In the oral health field, genome-wide association studies are not yet common, but several recent studies targeted at a handfull of candidate genes have yielded significant results. Nibali *et al*. [15] found associations between alleles at the interleukin IL-6 gene and aggressive periodontitis and then, in an effort to clarify the nature of the association, also found associations between SNPs in the IL-6 promoter region and the detection of pathogenic bacteria *A. actinomycetemcomitans* and *P. gingivalis* in subjects with severe periodontitis [43]. Palikhe *et al*. [44] also looked at the relationship between genetics, periodontitis, and the presence of

pathogeic bacteria in coronary artery disease patients using a logistic regression model that allowed them to include several covariates. They found that an allele (LTA+496C) at a locus within the major histocompatibility complex (MHP) on chromosome 6 is associated with periodontitis in patients with CAD.

These recent successes, as well as the observation that many oral health diseases have at least some genetic risk factors, suggest that genetic association studies may prove to be a valuable tool in the search for molecular basis underlying oral health disease.

References

[1] A. O. M. Wilkie & G. M. Morriss-Kay (2001) Genetics of craniofacial develoopment and malformation. *Nature Reviews Genetics*, **2**: 458–68.

[2] J. Dixon, S. J. Edwards, A. J. Gladwin *et al*. (1996) Positional cloning of a gene involved in the pathogenesis of treacher collins syndrome. *Nature Genetics*, **12**: 130–6.

[3] D Poswillo (1973) The pathogenesis of the first and second branchial arch syndrome. *Oral Surg. Oral Med. Oral Pathol.*, **35**: 302–28.

[4] H. Naora, M. Kimura, H. Otani, *et al*. (1994) Transgenic mouse model of hemifacial microsomia: cloning and characterization of insertional mutation region on chromosome 10. *Genomics*, **23**: 515–19.

[5] D. Kelberman, J. Tyson, D. C. Chandler, *et al*. (2001) Hemifacial microsomia: progress in understanding the genetic basis of a complex malformation syndrome. *Human Genetics*, **109**: 638–45.

[6] B. R. Rollnick & C. I. Kaye (1983) Hemifacial microsomia and variants: pedigree data. *Am. J. Med. Genet.*, **15**: 233–53.

[7] T. C. Hart (1996) Genetic risk factors for early-onset periodontitis. *Journal of Periodontology*, **67**: 355–66. Suppl. S.

[8] B. S. Michalowicz (1994) Genetic and heritable risk-factors in periodontal disease. *Journal of Periodontology*, **65**: 479–88. Suppl. S.

[9] A. Berdeli, G. Emingil, A. Gurkan, G Atilla, & T Kose (2006) Association of the IL-1RN2 allele with periodontal diseases. *Clinical Biochemistry*, **39**: 357–62.

[10] M. P. Cullinan, B. Westerman, S. M. Hamlet, *et al*. (2008) Progression of periodontal disease and interleukin-10 gene polymorphism. *Journal of Periodontal Research*, **43**: 328–33.

[11] G. Emingil, A. Berdeli, H. Baylas, *et al*. (2007) Toll-like receptor 2 and 4 gene polymorphisms in generalized aggressive periodontitis. *Journal of Periodontology*, **78**: 1968–77.

[12] Y.-P. Ho, Y.-C. Lin, Y.-H. Yang, K.-Y. Ho, Y.-M. Wu, & C.-C. Tsai (2008) Cyclooxygenase-2 gene(-765) single nucleotide polymorphism as a protective factor against periodontitis in taiwanese. *Journal of Clinical Periodontology*, **35**: 1–8.

[13] J. A. James, K. V. Poulton, S. E. Haworth, *et al*. (2007) Polymorphisms of tlr4 but not cd14 are associated with a decreased risk of aggressive periodontitis. *Journal of Clinical Periodontology*, **34**: 111–17.

[14] K. S. Kornman, A. Crane, H. Y. Wang, *et al*. (1997) The interleukin-1 genotype as a severity factor in adult periodontal disease. *Journal of Clinical Periodontology*, **24**: 72–7.

[15] L. Nibali, G. S. Griffiths, N. Donos, *et al*. (2008) Association between interleukin-6 promoter haplotypes and aggressive periodontitis. *Journal of Clinical Periodontology*, **35**: 193–8.

[16] A. P. Sumer, N. Kara, G. C. Keles, S. Gunes, H. Koprulu, & H. Bagci (2007) Association of interleukin-10 gene polymorphisms with severe generalized chronic periodontitis. *Journal of Periodontology*, **78**: 493–7.

[17] T. Tervonen, T. Raunio, M. Knuuttila, & R. Karttunen (2007) Polymorphisms in the CD14 and IL-6 genes associated with periodontal disease. *Journal of Clinical Periodontology*, **34**: 377–83.

[18] S. Li, M. H. Yang, C. A. Zeng, *et al*. (2008) Association of vitamin d receptor gene polymorphisms in chinese patients with generalized aggressive periodontitis. *Journal of Periodontal Research*, **43**: 360–3.

[19] R. B. de Brito, R. M. S. Scarel-Caminaga, P. C. Trevilatto, A. P. de Souza, & S. P. Barros (2004). Polymorphisms in the vitamin d receptor gene are associated with periodontal disease. *Journal of Periodontology*, **75**: 1090–5.

[20] K. Inagaki, E. A. Krall, J. C. Fleet, & R. I. Garcia (2003) Vitamin d receptor alleles, periodontal disease progression, and tooth loss in the va dental longitudinal study. *Journal of Periodontology*, **74**: 161–7.

[21] P. Concolino, F. Cecchetti, C. D'Autilla, *et al*. (2007) Association of periodontitis with GSTMI/GSTTI-null variants - a pilot study. *Clinical Biochemistry*, **40**: 939–45.

[22] H. C. Hung, J. Chuang, Y. C. Chien, *et al*. (1997) Genetic polymorphisms of cyp2e1, gstm1, and gstt1; environmental factors and risk of oral cancer. *Cancer Epidemiology Biomarkers and Prevention*, **6**: 901–5.

[23] F. M. Chung, Y. H. Yang, C. H. Chen, C. C. Lin, & T. Y. Shieh (2005) Angiotensin-converting enzyme gene insertion/deletion polymorphism is associated with risk of oral precancerous lesion in betel quid chewers. *British Journal of Cancer*, **93**: 602–6.

[24] E. Vairaktaris, C. Yapijakis, C. Tsigris, *et al*. (2007) Association of angiotensin-converting enzyme gene insertion/deletion polymorphism with increased risk for oral cancer. *Acta Oncologica*, **46**: 1097–1102.

[25] Y. Kuroda, H. Nakao, K. Ikemura, & T. Katoh (2007) Association between the tp53 codon72 polymorphism and oral cancer risk and prognosis. *Oral Oncology*, **43**: 1043–48.

[26] J. R. Gonzales, J. Michel, A. Diete, J. M. Hermann, R. H. Bodeker, & J. Meyle (2002) Analysis of genetic polymorphisms at the interleukin-10 loci in aggressive and chronic periodontitis. *Journal of Clinical Periodontology*, **29**: 816–22.

[27] R. C. Lewontin (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, **49**: 49–67.

[28] W. G. Hill & B. S. Weir (1994). Maximum likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics*, **54**: 705–14.

[29] M. Lynch & B. Walsh (1998). Genetics and Analysis of Quantitative Traits. *Sinaur Associates, Inc.*.

[30] B. S. Weir, L. R. Cardon, A. D. Anderson, D. M. Nielsen, & W. G. Hill. (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, **15**: 1468–76.

[31] The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**: 1299–1320.

[32] Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**: 661–78.

[33] P. D. Sasieni (1997) From genotypes to genes: Doubling the sample size. *Biometrics*, **53**: 1253–61.

[34] A. Agresti (2002) *Categorical Data Analysis*. John Wiley & Sons, Ltd, second edition.

[35] C. D. Campbell, E. L. Ogburn, K. L. Lunetta, *et al*. (2005) Demonstrating stratification in a European American population. *Nature Genetics*, **37**: 868–72.

[36] W. C. Knowler, R. C. Williams, D. J. Pettitt, & A. G. Steinberg (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in american indians with genetic admixture. *American Journal of Human Genetics*, **43**: 520–6.

[37] B. Devlin & K. Roeder (1999) Genomic conrol for association studies. *Biometrics*, **55**: 997–1004.

[38] D. Shmulewitza, J. Zhanga, & D. A. Greengerg (2004) Case-control association studies in mixed populations: Correcting using genomic control. *Human Heredity*, **35**: 235–54,.

[39] N. J. Samani, J. Erdmann, A. S. Hall, *et al*. (2007) Genomewide association analysis of coronary artery disease. *New England Jounal of Medicine*, **357**: 443–53.

[40] J. Hampe, A. Franke, P. Rosenstiel, *et al*. (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature Genetics*, **39**: 207–11.

[41] A. Zhernakova, A. Z. Behrooz, M. Bevova, *et al*. (2007) Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of a type 1 diabetes point to a general risk locus for autoimmune diseases. *American Journal of Human Genetics*, **81**: 1284–88.

[42] M. G. Hayes, A. Pluzhnikov, K. Miyake, *et al*. (2007) Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes*, **56**: 3033–44.

[43] L. Nibali, M. S. Tonetti, D. Ready, *et al*. (2008) Interleukin-6 polymorphisms are associated with pathogenic bacteria in subjects with periodontitis. *Journal of Periodontology*, **79**: 677–83.

[44] A. Palikhe, M.-L. Lokki, P. J. Pussinen, *et al*. (2008) Lymphotoxin alpha LTA+496C allele is a risk factor for periodontitis in patients with coronary artery disease. *Tissue Antigens*, **71**: 530–7.

18

The Bayesian approach

Emmanuel Lesaffre, Arnošt Komárek and Alejandro Jara

18.1 Introduction

In Chapter 10 we reviewed the classical statistical approaches to inference, where 'classical' refers to the hypothesis testing approach of Fisher and that of Neyman and Pearson. Although both approaches seem to combine quite well, the originators strongly disagreed in their views on how extrapolation from observed data to the population of interest should be done. In next section we elaborate briefly on this difference in viewpoints and review some of the difficulties with classical statistical inference.

Research is not done in isolation. When a new study is planned, say comparing a new treatment for treating oral cancer with a standard treatment, almost invariably a lot of background information is available on the two treatments. This information is incorporated to write the protocol of the trial, but is not explicitly used in a classical statistical analysis of the trial results. Suppose that you have set up a small sized RCT and that it shows an unexpectedly positive result about the a new treatment. Your first reaction (certainly of your drug company) would probably be 'great!' However, if in the past none of the drugs had such a large effect and biologically the new drug is not much different from the standard drug then probably you would become more cautious (if you are honest of course). By doing so you combine your prior knowledge with the observed results. There exists a quantitative way to incorporate the acquired knowledge into the trial result, i.e. using the Bayesian approach. This is the topic of this chapter.

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

We will show that the Bayesian approach differs from the classical approach in that (1) it is based only on the observed data, (2) external information can be incorporated into the analysis and (3) that the computational procedures are quite different.

We start with reviewing the classical frequentist way of statistical inference.

18.2 A reflection on classical statistical approaches

The goal of empirical research is to draw conclusions from the experiment (observed data) to the future, the theory, etc. Statistical inference aims at helping the researcher in this process. Both Fisher (1925, 1935) and Neyman and Pearson (1933) have greatly contributed to statistical inference, but while Fisher was aiming at inductive reasoning, Neyman and Pearson were strong proponents of deductive reasoning. This dramatic difference between the two approaches is not realized by many researchers, not even by some statisticians.

To fix ideas, the following illustration is taken from the Signal-Tandmobiel® (ST) study (see Chapter 19 for details). Among the seven-year-old children examined in the first year of the ST study, 56 % ($p = 0.56$) demonstrated caries experience (CE), whereby CE is defined as a binary variable indicating whether at least one of the deciduous teeth in the mouth is decayed (at d_3 level), missing or filled due to caries. Inspired by some data explorations and results from the literature, the oral health researchers wished to confirm in a statistical sense the observed East–West gradient in CE: for West Flanders 49.7 % compared to 63.2 % in Limburg (east of Flanders). A chi-squared test was performed and a test statistic $X^2 = 28.05$ obtained yielding a P-value of $P < 0.0001$. The P-value shows, according to Fisher, the evidence against the null hypothesis H_0. Using a threshold like 0.05 to claim evidence against H_0 was called a significance test by Fisher. There is, therefore, strong evidence that the two regions differ in CE experience. It must be stressed, however, that even without any data we would not really believe that H_0 can be true since it is hard to believe that exact equality of CE in two regions can ever exist in practice. Note also that the aim of Fisher was to provide the researcher a tool for inductive reasoning by introducing the P-value. As seen in Chapter 10, the calculation of the P-value is based on a frequency concept. That is, the P-value is a limiting ratio of fictive experiments done under H_0.

In contrast, Neyman and Pearson did not believe that a single experiment could provide valuable evidence against an hypothesis. Instead, they proposed the hypothesis test, a decision rule that makes the choice between H_0 and H_A (alternative hypothesis). The mechanism of such an hypothesis test has been explained in Chapter 10 and implies the definition of a (5 %) critical region, e.g. $X^2 > 3.84$ for the $\chi^2(1)$-test. If a result fell into the critical region, then H_A was to be accepted and H_0 rejected, otherwise H_0 was to be accepted. For our example the test statistic is X^2 and the observed value exceeds the 5 % threshold value (3.84) of the $\chi^2(1)$-distribution. Therefore the null-hypothesis is rejected here. Note in Neyman & Pearson's approach H_0 can be 'accepted' in contrast to what is being

taught nowadays in every course on medical statistics. Neyman & Pearson's rule is based on the repeated sampling idea; it has the property that in the long run H_0 is incorrectly rejected in 5 % of the times (i.e. when H_0 is true) for $\alpha = 0.05$. The decision rule is an example of deductive probability theory. Indeed, Neyman and Pearson claimed that the best one can do is to set up a rule that works well in the long run. Thus, also Neyman and Pearson need a frequency concept. Therefore classical statistical theory is also referred to frequentist statistics.

The two views are, however, nowadays combined into a single theory of statistical inference, despite the strong disagreement of their originators. For instance, results of a significance test are reported either by the obtained P-value (Fisher's view) or using the notation of '*' when $P < 0.05$ (Neyman & Pearson). Some statisticians are quite critical on the current tradition of melting the two approaches into a unified theory, see e.g. Goodman (1993), Goodman (1999a), Lilford and Braunholtz (1996) and Hubbard and Bayarri (2003).

The leading medical journals recognize nowadays that reporting solely a P-value is most often clinically uninformative. Further, the P-value is not always clearly understood by applied researchers and therefore misuses of the P-value are abundant in the medical/OH literature. For this reason, the confidence interval is preferred as a tool for (inductive) statistical inference. It is, however, debatable that the confidence interval can meet its expectations. Indeed the 95 % CI is defined as a random interval that encompasses the true value of the parameter in (approximately) 95 % of (repeated) studies. Thus, the confidence interval gets its meaning only when we imagine that the study will be repeated (indefinitely). But, for a single experiment the 95 % CI either includes or excludes the true value. Yet, in practice the meaning of the 95 % confidence interval is formulated and interpreted differently. For our example, the observed difference in CE (Limburg – West Flanders) is equal to 13.5 % with a 95 % confidence interval of [5.9 %, 21.1 %]. It is customary to state that with 95 % chance the true difference in CE lies between 5.9 % and 21.1 %. However, the covering property of the confidence interval is a property of the family and not of the computed interval. Indeed, we could consider the interval [5.9 %, 21.1 %] as a random draw from a population of intervals that includes or covers the true value of the parameter 95 % of the times. Therefore, the logic behind the use of 95 % confidence is that the computed interval has a high chance of being one of the good ones that covers the truth. But this interpretation differs from how clinicians most often think about the confidence interval.

In spite of the above critical remarks we do not wish to condemn the classical statistical approaches. In fact, frequentist methods have been quite successful in taking into account the uncertainty associated with empirical results into the interpretation of empirical results. However, this does not imply that they constitute the only way of interpreting experimental results or that they are necessarily best. Nevertheless, it appears that a tool that is only based on the observed data and that allows direct (inductive) inference from the data to the population is most welcome. This is provided with the likelihood function and further developed into the Bayesian approach whereby it is also possible to incorporate external information.

18.3 Bayes' Theorem: the basics

Bayes' Theorem is a mathematical result that constitutes the fundament for the Bayesian approach and is attributed to Thomas Bayes, born around 1701 and died in 1761. His (now) famous theorem was submitted posthumously by his friend Richard Price in 1763 to *Philosophical Transactions of the Royal Society* and was entitled *An Essay toward a Problem in the Doctrine of Chances*. It took, however, a few centuries to appreciate the impact of this theorem on applied research Bayes (1763). The basic version of Bayes' Theorem is explained first using the above example on the prevalence of CE in Flanders.

Table 18.1 shows the observed frequencies of 7-year-old children with (D = 1) or without CE (D = 0) included in the ST study, who are either from West Flanders (Y = 0) or Limburg (Y = 1). Imagine that three OH researchers are interested in estimating the probability that a seven-year-old child from Limburg demonstrates CE.

The first OH researcher randomly draws a child from the total group of children. Assuming that the observed proportions are in fact probabilities, the probability that a child from Limburg has CE can then be read off directly from Table 18.1, i.e. $477/1512 = 0.32$. This probability is called the joint probability of observing $Y = 1$ and $D = 1$ and is denoted $Pr(Y = 1, D = 1)$. The second OH researcher samples the children in two stages. In the first stage, he samples children according to the geographical location and then whether the child has CE or not. The probability that in the first stage the child belongs to Limburg is $Pr(Y = 1)$, here equal to $755/1512 = 0.50$. The conditional (on living in Limburg) probability that a child with CE is selected, $Pr(D = 1|Y = 1)$, is here equal to $477/755 = 0.63$. The third OH researcher also decides to sample in two stages, but in the reverse order. The first probability (of selecting a child with CE) happens with probability $Pr(D = 1) = 853/1512 = 0.56$ and then a child is sampled from Limburg with probability $Pr(Y = 1|D = 1) = 477/853 = 0.56$. Clearly, all three OH researchers must end up with the same probability of sampling a child from Limburg with CE. Indeed, it can be easily checked that

$$Pr(Y = 1) \times Pr(D = 1|Y = 1) = Pr(Y = 1, D = 1)$$
$$= Pr(D = 1) \times Pr(Y = 1|D = 1),$$

Table 18.1 Signal-Tandmobiel® study: seven-year-old children living either in West Flanders or Limburg split up according to caries experience (C = caries, NC = no caries).

	Caries experience		
Province	NC (D = 0)	C (D = 1)	Total
West Flanders (Y = 0)	381	376 (49.7 %)	757
Limburg (Y = 1)	278	477 (63.2 %)	755
Total	659	853	1512

since $0.50 \times 0.63 = 0.32 = 0.56 \times 0.56$. Note that the joint probability $Pr(Y = 1, D = 1)$ is different from the two conditional probabilities.

The above expression leads to the famous Bayes' Theorem given by:

$$Pr(D = 1|Y = 1) = \frac{Pr(Y = 1|D = 1)Pr(D = 1)}{Pr(Y = 1)}. \tag{18.1}$$

Bayes' Theorem is the basis for calculating the predictive values from the sensitivity and specificity of a diagnostic test. To see this, take Example 12.1. Let oral mucosal lesions patients with (without) dysplasia or carcinoma denoted by $D = 1$ ($D = 0$), then $Pr(D = 1)$ is the (true) prevalence of the disease. Further, let a positive (negative) diagnostic test OralCDx be indicated by $Y = 1$ ($Y = 0$) then $Pr(Y = 1)$ is called the apparent prevalence of the disease. With Bayes' Theorem one can calculate the conditional probability of being diseased given a positive test result, $Pr(D = 1|Y = 1)$, called the positive predictive value (PPV), from the reversed conditional probability $Pr(Y = 1|D = 1)$ (sensitivity) and the true and apparent prevalence. In a similar manner the negative predictive value (NPV) is obtained.

According to a general property in probability theory (Law of Total Probability) the apparent prevalence can be written as: $Pr(Y = 1) = Pr(Y = 1|D = 1)Pr(D = 1) + Pr(Y = 1|D = 0)Pr(D = 0)$. This equation expresses the probability of a positive test as the sum of two probabilities: (1) the probability of showing correctly a positive test times the probability that the subject is diseased and (2) the probability of showing falsely a positive result times the probability that the subject is healthy. Using this result, expression (18.1) can be rewritten as:

$$Pr(D = 1|Y = 1) = \frac{Pr(Y = 1|D = 1)Pr(D = 1)}{Pr(Y = 1|D = 1)Pr(D = 1) + Pr(Y = 0|D = 0)Pr(D = 0)}, \tag{18.2}$$

which expresses PPV as a function of the prevalence, the sensitivity and the specificity of the test, see also Section 12.1. An equation similar to (18.2) leads to the expression for NPV.

Both examples illustrate Bayes' Theorem in a classical frequentist context, but they are not sufficient to illustrate adequately the Bayesian approach. Let us therefore take again the previous example, and suppose that a dentist wishes to predict whether his/her patient has dysplasia or carcinoma if (s)he shows oral mucosal lesions. Prior to knowing the result of the OralCDx test the dentist can only state that the probability that the patient is diseased ($Pr(D = 1)$) is equal to the prevalence of the disease in the population ($26/96 = 0.27$). Upon receiving the information that the OralCDx test is positive, the dentist will adjust his/her opinion and arrives at a probability of 0.89 which is in fact $Pr(D = 1|Y = 1)$ (= PPV). In this context $Pr(D = 1)$ is called the prior probability while the test result ($Y = 1$) represents the data (corresponding to the likelihood, see below). The probability $Pr(D = 1|Y = 1)$ is obtained from combining the data with the prior probability and is called the posterior probability. When the prevalence for the disease is not known, the dentist could still have a prior probability simply by guessing or from his/her expert knowledge. In that case we speak of a prior belief.

In conclusion, patients with oral mucosal lesions must have either dysplasia or carcinoma or not but their true status is unknown to the dentist. The dentist can/will have an opinion about the disease status of the patient and can then use Bayes' Theorem to combine his/her prior probability/belief $Pr(D = 1)$ with the information obtained from the data (the likelihood $Pr(Y = 1|D = 1)$) to arrive at a revised probability of suffering from the disease, i.e. the posterior probability $Pr(D = 1|Y = 1)$.

18.4 Bayes' Theorem: the general rule

In the previous section Bayes' Theorem was used to update our inference on an individual subject. Here we are interested in drawing inference on a population, which will lead us to the general Bayes' Theorem (also called Bayes' rule).

Suppose that, in contrast to previous section, we are not sure about the prevalence π of CE in Flanders and that we are doubting between several, though reasonable, values for π. Suppose for now that the possible values for π are 0.50, 0.60 and 0.70 and that historical information points to a prevalence of most likely 0.60. A prior probability of 0.70 was attached to $\pi = 0.60$, but the other two values for the prevalence were also considered possible, though less likely. The prior probabilities were 0.20 for $\pi = 0.50$ and 0.10 for $\pi = 0.70$. Subsequently, an oral health survey conducted in Flanders yielded 25 out of 50 seven year old children with CE on deciduous teeth. In a classical statistical analysis one would rely only on the observed proportion of 0.50 and neglect the historical information. In the Bayesian approach the observed proportion is combined with the historical information to arrive at a posterior statement. This will now be shown.

The information that the dental survey provides about π is expressed by the binomial likelihood function, introduced in Chapter 10. The probability that 25 out of 50 children show CE if the true probability is π is given by

$$L(\pi|Y = 25) = \binom{50}{25}\pi^{25}(1 - \pi)^{25}. \tag{18.3}$$

As a function of π, $L(\pi|Y = 25)$ is called the binomial likelihood function and expresses the plausibility of the observed result $Y = 25$ (for a sample of size $n = 50$) for different potential values of π. In the present (artificial) setting, only 3 values of π are of interest and hence there are only three values for the likelihood function: $L(0.50|Y = 25) = 0.11$, $L(0.60|Y = 25) = 0.04$ and $L(0.70|Y = 25) = 0.001$. Clearly, the survey supports most the value $\pi = 0.50$ and in a classical, frequentist analysis we would decide that $\pi = 0.50$ is the most plausible value. One must realize, though, that another survey will probably give another proportion supporting another value for π. Also, if there is strong prior evidence for $\pi = 0.60$ then why ignoring it. This brings us to the question 'Which value do we prefer for π in the light of the prior knowledge and the observed result (likelihood)?' Clearly, values that are supported both by the data and the prior belief should be

preferable leading to the more general expression of Bayes' Theorem:

$$p(\pi_i|y) = \frac{L(\pi_i|y)\,p(\pi_i)}{\sum_{j=1}^{3} L(\pi_j|y)\,p(\pi_j)},\ (i = 1, 2, 3) \tag{18.4}$$

where y represents here that $Y = 25$ was obtained and the only possible values for the prevalence are here $\pi_1 = 0.50$, $\pi_2 = 0.60$ and $\pi_3 = 0.70$. Equation (18.4) gives the posterior probability for π and the RHS of this equation shows that when the likelihood and the prior probability for a particular π is high, then the posterior probability is also high. The denominator ensures that the posterior probabilities add up to one for the three values of π. The posterior probability of $\pi = 0.50$, 0.60 and 0.70 is equal to 0.441, 0.556 and 0.003, respectively. From these probabilities one can infer that the highest posterior belief is attached to $\pi = 0.60$, which is the same as the value one postulated at the start. One should not conclude, though, that the collected data had no influence on our decision. Indeed, our prior belief for $\pi = 0.50$ was 0.20, while the posterior belief increased to 0.441. The reason for not choosing 0.50 is that the information contained in the sample was not strong enough to overrule the prior belief. With a different result on the survey or the same result with a larger survey the conclusion might be quite different.

In practice π can vary freely between 0 and 1. Suppose therefore that experts believe that the most likely value for the prevalence of CE in Flanders is 0.60, but that there is uncertainty and claim that with high probability π lies between 0.50 and 0.70. This could be reformulated in a prior belief that π belongs to [0.50, 0.70] with 0.95 probability. Hence π is not considered as a fixed value anymore as in classical statistics, instead it is treated as a random variable with a probability distribution $p(\pi)$ as in Figure 18.1. This figure expresses our uncertainty about π. Namely, the highest value for $p(\pi)$ occurs for $\pi = 0.60$, which is called the prior mode and is the most plausible a priori value for π. Further, the area under the curve between 0.50 and 0.70 is 0.95, and corresponds to the prior belief of the experts stated above.

To combine prior knowledge and the information obtained from the data we need a further generalization of Equation (18.4). For a continuous π, the likelihood is still given by the binomial likelihood (18.3) but now it is a continuous function of π (on [0,1]). Figure 18.2 shows that for $Y = 25$ the likelihood function is quite low for values of π close to 0 and 1, but relatively high between 0.3 and 0.6. The maximum value of the likelihood function, the maximum likelihood estimate (MLE) is 0.5, also equal to the observed proportion. The mathematical combination of the prior distribution with the likelihood leads to the following Bayes' rule for continuous parameters:

$$p(\pi|y) = \frac{L(\pi|y)\,p(\pi)}{\int L(\pi|y)\,p(\pi)\,d\pi} = \frac{L(\pi|y)\,p(\pi)}{p(y)}. \tag{18.5}$$

The denominator is now an integral (loosely speaking the limit of a sum when the number of possible values for π goes to infinity). In Figure 18.2 the posterior

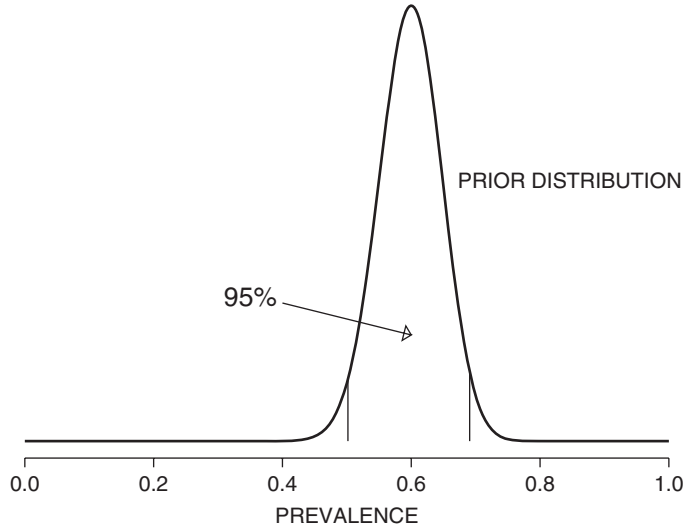


Figure 18.1 Prevalence of caries experience: Prior distribution on prevalence, the 95 % prior boundaries are indicated by vertical lines.

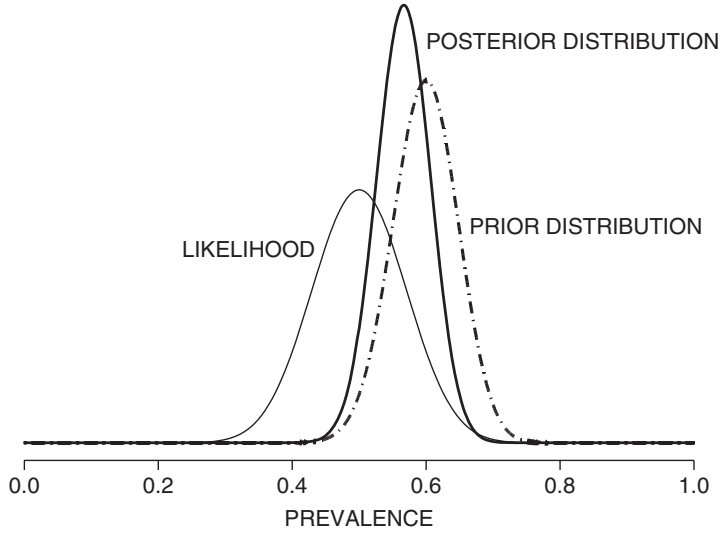


Figure 18.2 Prevalence of caries experience: Prior- and posterior distribution together with binomial likelihood on prevalence.

distribution $p(\pi|y)$ is shown for our example together with the prior distribution and expresses what values of π are plausible in the light of the study and the prior information. The most plausible posterior value, the posterior mode, is the value of π that maximizes $p(\pi|y)$ and is here equal to 0.57. Further, the a posteriori

uncertainty about π can be expressed as 'with 0.95 probability π lies between 0.49 and 0.64' since the area under $p(\pi|y)$ from 0.49 and 0.64 equals 0.95. All of this can be checked graphically on Figure 18.2. Further, it can be seen that the posterior distribution lies somewhat in-between the prior distribution and the likelihood function and similarly the posterior mode lies in-between the prior mode and the MLE. Thus, to arrive at the posterior information a compromise is made between the prior information and the information obtained from the sample. Finally, the posterior distribution is more peaked than the prior distribution, showing that there is more information about π in $p(\pi|y)$ than in $p(\pi)$.

The denominator in (18.5) does not depend on π. Thus, Bayes' rule transforms prior information into posterior information as follows:

$$\text{posterior} \propto \text{likelihood} \times \text{prior},$$

where $\propto$ means 'proportional to' and the prior and posterior distribution are abbreviated as *prior* and *posterior*, respectively. It is important to remember that in the Bayesian paradigm a parameter is treated as a random variable and its probability is an expression of the belief we have about that parameter.

18.5 Extracting information from the posterior

The posterior contains all information about the parameter of interest. For instance, the posterior probability that the prevalence of CE is in-between 0.5 and 0.7 can be obtained from the integral

$$\int_{0.5}^{0.7} p(\pi|y)\, d\pi.$$

For the posterior in Figure 18.3 the probability that π is more than 0.3 is equal to 0.06, indicated on the figure. For a posterior corresponding to a classical distribution, such as the Gaussian distribution, this integral is readily obtained from standard software while for a non-classical posterior often dedicated software needs to be written.

To ease the communication, the posterior needs to be characterized by summary measures, just as in classical statistics the sample is characterized by the average, standard deviation, etc. The posterior mode $\hat{\pi}_M$ is one possible summary measure. Because $p(\pi|y)$ is a distribution we can also determine its mean, median, standard deviation, etc. requiring, however, integral calculation. For instance, the posterior mean of π is defined as

$$\bar{\pi} = \int_0^1 \pi\ p(\pi|y)\, d\pi, \tag{18.6}$$

as in expression (10.3) (Chapter 10). The posterior median $\bar{\pi}_M$, is defined as the value of π such that the AUC (area under the curve) left to it is equal to 0.5. For the posterior in Figure 18.3, $\hat{\pi}_M = 0.167$, $\bar{\pi} = 0.188$ and $\bar{\pi}_M = 0.181$. The three

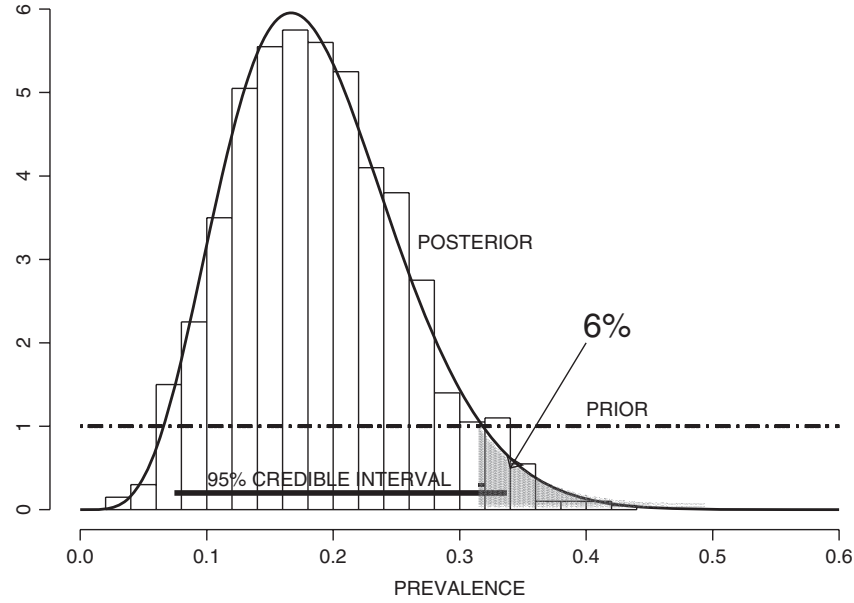


Figure 18.3 Prevalence of caries experience: Posterior obtained from combining a non-informative prior with the likelihood obtained from a survey in which five out of 30 children had caries experience. The overlaying histogram is obtained from 1000 sampled values from the posterior.

measures can be used to characterize the posterior, but a common choice is the posterior mean except when the posterior is skewed in which case the posterior median is preferred (as in our example). It is important to stress that the choice of the posterior point estimator is formally based on decision-theoretic arguments. The aforementioned summary measures only characterize the location of the posterior, but we also need to characterize the spread in $p(\pi|y)$. This is given by the posterior standard deviation $\bar{\sigma}$ derived from the posterior variance in a similar way as the posterior mean. For our example, $\bar{\sigma}$ is 0.068. The Bayesian equivalent of the classical (95 %) confidence interval is called the (95 %) credible interval (CI). A 95 % CI is the interval $[a, b]$ satisfying

$$P(a < \pi < b \mid y) = 0.95.$$

The above property does not uniquely define the credible interval, though. A popular version is the equal-tail 95 % CI. In this case a and b are chosen such that $P(\pi < a|y) = 0.025$ and $P(b < \pi|y) = 0.025$. This yields for our example (calculated using standard software) a 95 % equal-tail CI of $[0.075, 0.34]$, displayed in Figure 18.3. The interpretation of this interval is that 'we are uncertain about the true value of π but with 95 % (posterior) certainty we believe that the prevalence lies between 0.075 and 0.34'. Isn't this the way that we usually interpret our classical 95 % confidence interval?

18.6 Testing hypotheses in a Bayesian manner

We consider here two tools for hypothesis testing in the Bayesian context: (1) the Bayes' factor and (2) the contour probability.

To fix ideas let us take the following fictive split-mouth study. Two filling materials A and B are randomly assigned to two teeth in the mouth of $(n = 30)$ individuals. Suppose that for $(x = 21)$ patients filling material A remained longer intact than filling B. We analyze these data in a Bayesian manner in two settings of null- and alternative hypotheses. The parameter θ is the probability that filling material A is better than B, with $\theta = 0.5$ implying that A and B are equally good. The likelihood of the data is the binomial likelihood (18.3) with π replaced by θ and based on the observed data (21 successes for A out of 30 patients).

We consider the following two cases pertaining to θ: (a) $H_0 : \theta = 0.5$ and $H_a : \theta = 0.8$; (b) $H_0 : \theta = 0.5$ and $H_a : \theta \neq 0.5$. The posterior probability for choosing H_0 is in both cases:

$$p(H_0 \mid x) = \frac{p(x \mid H_0)\, p(H_0)}{p(x \mid H_0)\, p(H_0) + p(x \mid H_a)\, p(H_a)}. \tag{18.7}$$

To determine $p(H_0 \mid x)$ and $p(H_a \mid x)$ in our example, we calculate for case (a)

$$p(x = 21 \mid H_a) = \binom{30}{21} 0.8^{21}\, 0.2^9 = 0.0676,$$

$$p(x = 21 \mid H_0) = \binom{30}{21} 0.5^{21}\, 0.5^9 = 0.0133.$$

The ratio of the two posterior probabilities is given in general by

$$\frac{p(H_0 \mid x)}{p(H_a \mid x)} = \frac{p(x \mid H_0)}{p(x \mid H_a)} \times \frac{p(H_0)}{p(H_a)} \equiv BF(x) \times \frac{p(H_0)}{p(H_a)},$$

where $BF(x)$ is called the Bayes' factor contrasting the null hypothesis and the alternative hypothesis. In this case, the Bayes' factor corresponds to the likelihood ratio. In general it expresses how the prior odds (note that $p(H_a) = 1 - p(H_0)$) is transformed into the posterior odds. The Bayes' factor is therefore the index through which the data speak and is independent of the specified prior information. For the split-mouth study data, the Bayes' factor is equal to 0.197, hence there is preference for $\theta = 0.8$. Many Bayesians have a strong preference for the Bayes' factor as an inferential tool (see e.g. Goodman 1999b also for more background on the Bayes' factor).

Case (b) (most often used in practice) is a bit more tricky, since we need to compute

$$p(x = 21 \mid H_a) = \int_0^1 \binom{30}{21} 0.8^{21}\, 0.2^9\, p(\theta \mid H_a)\, d\theta,$$

where $p(\theta \mid H_a)$ is the prior distribution of the possible values of the parameter under the alternative hypothesis. This implies that in case (b) the Bayes' factor

depends on prior information. Goodman (1999b) argues that is not a weakness of the Bayes' factor but is a reflection of the ignorance of the researcher at the start. For the split-mouth study and with equal prior belief in H_0 and H_a, the Bayes' factor becomes 0.41 again showing preference for the alternative hypothesis. Compare this preference to the classical P-value of comparing H_0 versus H_a obtained from the likelihood ratio test (Chapter 10). We obtained $P = 0.026$ indicating a rather strong preference for H_a, while the preference as expressed by the Bayes' factor is not that extreme here.

The second Bayesian inferential tool is the contour probability, first suggested by Box & Tiao (1973). It is a kind of Bayesian P-value and measures the evidence against H_0 based on the posterior distribution of the parameter. Formally, it is one minus the posterior probability of the smallest credible interval (also called HPD-interval where the term 'HPD' is explained in e.g. Spiegelhalter *et al.* (2004) that just contains the value of θ under H_0. In Figure 18.3 the contour probability is indicated for $\beta_{gender} = 0$ (see below to understand the meaning of this parameter) by the area of the graded areas. For the split-mouth study, the contour probability for $H_0 : \theta = 0$ is equal to 0.023. Hence, the contour probability expresses the extremeness of the value of the parameter under H_0 in a similar way as a classical P-value except that now all inference is based on the observed data and the prior and is not based on a long run frequency concept.

18.7 Prior distributions

The Bayesian approach combines prior information with information extracted from an experiment to update our knowledge. In the previous prevalence example experts had a prior belief about the parameter π. The process of transforming the expert's knowledge about a parameter into a quantitative (read: 'probabilistic') language is called eliciting prior information (from experts). There are many ways to do this. In the prevalence example the expert was asked for the most likely value of π and for his/her (prior) uncertainty expressed e.g. as prior 25 % and 75 % quantiles. This is often not an easy task and experts often tend to disagree in their opinion. For the acceptance of the Bayesian analysis by the community the prior distribution should reflect the current opinion or current state of knowledge. This can be done by relying on a consensus prior representing the opinion of a group of experts, as was done in our prevalence example. For a review on issues involved in eliciting prior information, we refer to O'Hagan *et al.* (2007).

Prior information can also be extracted from previous experiments. In that case we speak of a data-based prior. An example of a prior determined from historical data is given in the example on emergence times below. A prior that represents information on the parameter, extracted either from prior knowledge or from prior studies, is called a subjective or informative prior.

While there is always some information available about a parameter, see e.g. Box and Tiao (1973), many of us would feel uncomfortable to express any preference for a particular value of that parameter in case we are ignorant about the

topic. A prior that corresponds to introducing no (subjective) information, is called a non-informative (NI) prior. Although it is not completely correct the uniform distribution is used in this context. A better name for this prior is a flat prior, shown in Figure 18.3 for the prevalence example. For this prior, the likelihood function is the same as the posterior and therefore one could claim that the prior is 'objective'. Many terms (vague, diffuse, reference, ...) are in use to indicate that the prior bears no or little prior information.

To establish the posterior distribution, the integral in the denominator of equation (18.5) needs to be calculated. In general this is a difficult task, but for some combinations of likelihood and prior the integral and hence the posterior is readily obtained. In the above prevalence example, we have taken as prior a beta distribution with parameters α_0 and β_0 (equal in the example to 61 and 41, respectively) given by

$$Beta(\alpha_0, \beta_0) = \frac{1}{B(\alpha_0, \beta_0)} \pi^{\alpha_0 - 1} (1 - \pi)^{\beta_0 - 1}, \tag{18.8}$$

with $B(\alpha_0, \beta_0)$ a value (the beta function) such that the AUC of the corresponding graph is one. This prior was a good representation of the prior belief expressed by the experts. When this prior is combined with a binomial likelihood, the posterior is again a beta distribution but with parameters $\bar{\alpha} = \alpha_0 + y$ and $\bar{\beta} = \beta_0 + n - y$ (in the prevalence example $\bar{\alpha} = 86$ and $\bar{\beta} = 66$). Since the posterior is of the same type as the prior, the beta distribution is called the conjugate prior (distribution) to the binomial likelihood.

Conjugate priors exist for a variety of models. For instance, a possible (but not good) model to describe the distribution of the dmft-index is a Poisson distribution. In that case, the conjugate prior is a gamma distribution which means that the posterior is again a gamma distribution. Another, quite important, case in statistics is the normal distribution which we will illustrate with an example on emergence times. The emergence age of a tooth is the chronological age of a child when that tooth appears in the mouth. Knowing the tooth emergence is of interest in dental practice for diagnosis and treatment planning. In forensic dentistry this information is useful for the estimation of the chronological age of a child with unknown birth records. In the ST study the emergence time (in years) of tooth 14 (FDI (Fédération Dentaire Internationale) notation) obtained from 2297 boys in the ST study is practically normally distributed with mean $\bar{y} = 10.7$ and standard deviation $s = 1.35$. The ST study was conducted in Flanders from 1996 to 2001. Suppose that one wishes to set up a similar but smaller-sized study in the Netherlands, or that in Flanders one has the intention to repeat the ST study within ten years but with fewer children. Suppose also that the only goal is to have a good estimate of the mean emergence time μ (quite unrealistic but it serves our purpose now). It can be shown that the data-based prior to using the ST data is a normal distribution with mean $\mu_0 = 10.7$ yrs and standard deviation $\sigma_0 = 1.35/\sqrt{2297} = 0.028$ yrs. This is quite precise prior information and perhaps the OH researchers want to downplay the impact of the prior (one can think of many reasons); this can be done by increasing σ_0 to an acceptable value. Or, the OH researchers feel that the

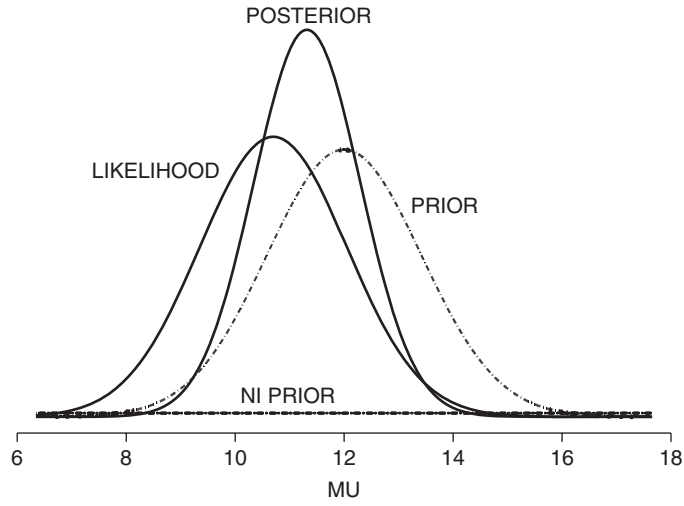


Figure 18.4 ST study: Normal prior, likelihood and posterior for emergence times of boys.

ST study is not completely representative for the current study. In that case both the prior mean and variance could be adapted. Suppose now that the OH researchers agree that a $N(10, 0.7^2)$ represents well the prior belief. The results of the smaller study can then be combined with the prior knowledge. When the observed mean emergence time is 11.5 yrs and the sample variance equal to $1.5\,\text{yrs}^2$ based on a sample of size 150, the posterior is again a normal distribution with mean 10.3 yrs with variance $0.63\,\text{yrs}^2$ (see Figure 18.4).

The emergence example points to the sequential nature of the Bayesian methodology. That is, the posterior from one experiment can be used as the prior for a subsequent experiment. However, combining 'objective' data with subjective (prior) information is not appreciated by many researchers. We need to realize, though, that research is never done in a purely objective manner. The routes that are taken in research do not only depend on objective results but also on the interpretation of these results. In the Bayesian approach this subjectivity is incorporated in a formal manner while in the classical, frequentist approach it is latent. For instance, clinical trialists acknowledge that it is quite unlikely to discover nowadays a wonder drug. A (highly) significant treatment effect from a small study will therefore be received with a great deal of skepticism. This is a Bayesian attitude since the trialists combine their prior (dis)belief with the results of the study to end up being much less enthusiastic, see e.g. Tan *et al.* (2003) for an interesting medical example. Finally, note that for a large sized study it does not matter too much how much prior information is specified, since the data (likelihood) will dominate the outcome and hence the prior information will hardly influence the posterior.

18.8 The Bayesian analysis of realistic examples

18.8.1 Multiparameter models

Most statistical models are based on more than one parameter (of interest). For instance, for the normal distribution both the mean and variance are parameters and might be of interest. In the above emergence example however, we assumed for reasons of simplicity that the variance of the data was known in advance (prior to taken the data). In a linear regression model, a logistic regression model, a Cox survival model, etc. (see Chapter 11) the set of parameters constitutes of the regression coefficients and possibly some extra model parameters, such as the error variance. When mixed effects models are employed (see Chapter 13), the random effects are treated in the same way as the model parameters parameters. In case of multiple parameters, prior information needs to be provided for all parameters to yield posterior distributions. While there is conceptually no difficulty to apply the Bayesian machinery to complex data models, the numerical difficulties to calculate the necessary integrals become unsurmountable. Several algorithms were suggested to replace formal integration, but only upon the introduction of advanced sampling techniques the Bayesian approach is able to tackle practical problems. Moreover, we show in Sections 18.8.2 and 18.8.3 that, with the recent computational advances, Bayesian methods are somewhat better suited to tackle complex statistical analyzes.

18.8.2 Markov chain Monte Carlo sampling

Formal calculation of the posterior (by mathematical derivations) can be replaced by sampling techniques. Sampling a real population was treated in Section 10.7.2. In the same chapter we have seen that a histogram obtained from a random sample is an estimate of the true distribution of the data in the population from which we can estimate the true mean, median, etc. In a Bayesian analysis we also need the mean, median, etc of the posterior distribution to characterize what we know about the parameter. In a similar way, but now using a computer program, one can sample from the (posterior) distribution. Indeed, the (posterior) distribution in Figure 18.3 could be viewed as a classical distribution representing a population. Using a sampling technique (in general called Monte Carlo sampling) one takes a random sample from such an (artificial) population yielding a histogram and summary measures. In Figure 18.3 we overlaid the mathematically determined posterior (beta distribution) with the histogram of 1000 sampled values from the same posterior. Clearly, the histogram approximates the posterior quite well. Also, the descriptive measures obtained from this sample approximate the true posterior measures closely.

Sampling techniques are essential in the Bayesian analysis of realistic examples, i.e. to determine summary measures from highly multivariate posterior distributions which are typical for dental problems. Indeed, with the hierarchical structure of

dental data, e.g. surfaces on teeth and teeth in a mouth, a large number of parameters are often involved. Further, in many dental problems one needs to model spatial correlations, take into account interval censoring, allow for misclassification, etc.. All of these features arise also in other medical problems, but in dentistry these problems come together. The Markov Chain Monte Carlo (MCMC) techniques are a class of techniques which replace the intractable integration methods by sampling. MCMC techniques have radically changed the way Bayesian analyzes are done nowadays and made the Bayesian approach quite popular. There are two types of MCMC sampling techniques: (a) Gibbs sampling and (b) Metropolis-(Hastings) sampling. The first was developed by Geman and Geman (1984) and introduced in statistics by Gelfand and Smith (1990). The second type was introduced by Metropolis *et al*. (1953) and further developed by Hasting (1970). The mechanism of MCMC sampling differs greatly from the simple random sampling technique that was used in Figure 18.3, but it goes beyond the scope of this chapter to elaborate on its technical background. It suffices to say here that MCMC techniques generate a sequence of observations from the posterior distribution, so called a Markov chain, using a Monte Carlo technique (this explains the term MCMC, i.e. Markov Chain Monte Carlo) from which the posterior summary measures can be determined. Technical assistance from a statistician is an absolute must, since it needs a careful check that the sampling program is producing the correct posterior distribution. Indeed the program needs to have converged before it can produce the correct summary measures. In other words, a Bayesian analysis of a realistic data analysis requires considerable technical skills of the user. But, in return the MCMC techniques allow tackling virtually any complex data problem in a Bayesian manner. The development of the statistical MCMC software WinBugs, see http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml, has greatly contributed to the popularity of the Bayesian approach. This software is based on Gibbs Sampling, one of the two MCMC techniques and requires minimal programming efforts even when complex statistical models are required to analyze the data.

The medical and oral health literature finds an increasing number of papers making use of the Bayesian approach. In general, the reasons for the researcher to switch to the Bayesian paradigm are either (1) the researcher prefers the Bayesian way of statistical inference and/or (2) the statistical approach to tackle the research questions is too complex to handle it in a classical frequentist context and therefore needs MCMC techniques. Thus, in the second case the motivation for acting Bayesian is merely pragmatic. We must admit that many applied statisticians choose nowadays for a Bayesian approach primarily for its computational flexibility, and less for its philosophical roots.

In next two sections we illustrate the use of MCMC techniques in complex problems. In principle one might analyze these research questions also in a frequentist manner, but the Bayesian MCMC techniques allow to try out (in a flexible manner) many (complex) statistical models on the data. On the other hand, if one fundamentally believes in the Bayesian paradigm then this is probably the only way to tackle the research questions. Further, we will indicate at several places the attractiveness of the Bayesian approach.

18.8.3 Two complex analyses using MCMC techniques

18.8.3.1 Modeling the risk of caries experience

The data and the research question The *Smile for Life (SFL)* study is an ongoing oral health promotion intervention study, which involves two groups of newborns (with and without the intervention) receiving an extensive oral examination at three and at five years of age. At the planning stage of the study, three- and five-year-old children from the four geographical areas in Flanders that were to be involved in the SFL study, underwent an oral examination (pilot study). Here the data of the five-year-old children will be analyzed. Further details of the SFL study can be found in Chapter 19 and in Declerck *et al.* (2008).

One research aim in the pilot (but also in the intervention programme) is to assess the association of demographic, dietary and brushing behavior variables with the presence of CE in primary dentition. The CE status was assessed by seven calibrated and trained dentists, see Chapter 19 for more details on the calibration sessions. However, despite the calibration exercises, the dentists still showed a different classification behavior compared to a benchmark scorer. Thus, the true CE is not known for the children involved in the pilot study and the same is/will be true for the children involved in the intervention programme. This complicates the identification of risk factors. In order to establish the impact of the risk factors for CE, one needs to correct for misclassification (see Chapter 16).

Let the variable D_{ij} denote the true CE state of the jth, $j = 1, \ldots, n_i$ child examined by the ith dentist, $i = 1, \ldots, 7$, namely, $D_{ij} = 1$ indicates that CE is truly present and $D_{ij} = 0$ indicates that it is absent. Further, let $Y_{ij} = 1$ indicate that the ith dentist classifies the jth child as demonstrating CE and $Y_{ij} = 0$ indicates otherwise. The sensitivity and specificity for the ith dentist are $\alpha_i = Pr(Y_{ij} = 1 \mid D_{ij} = 1)$, and $\eta_i = Pr(Y_{ij} = 0 \mid D_{ij} = 0)$, respectively. To examine the association Declerck *et al.* (2008) used a logistic regression model where the misclassification process was taken into account, assuming that all α_i and η_i (used to correct for misclassification) are known. Thus, this analysis ignored that the parameters α_i and η_i were in fact estimated. The impact of this is that the variability of the estimated logistic regression parameters is underestimated. Here we illustrate the Bayesian learning process by using the data obtained after the calibration exercise to create a prior distribution on the misclassification parameters.

The model Let the true probability for CE be equal to $\pi(\boldsymbol{x}_{ij}) = Pr(D_{ij} = 1 \mid \boldsymbol{x}_{ij}, \boldsymbol{\beta})$, with $\boldsymbol{x}_{ij}$ a vector of covariates and $\boldsymbol{\beta}$ a vector of regression coefficients. A popular way to describe the risk of CE is done with logistic regression (see Chapter 11) given by $\pi(\boldsymbol{x}_{ij}) = \exp\left\{\boldsymbol{\beta}^T \boldsymbol{x}_{ij}\right\} / \left(1 + \exp\left\{\boldsymbol{\beta}^T \boldsymbol{x}_{ij}\right\}\right)$. Note that we contrasted the logistic regression model to other models but found no reason to doubt the chosen model.

We do not necessarily observe the true CE status (D_{ij}) of the child, but a possibly misclassified status (Y_{ij}). The model that relates the observed response and covariates is based out of the model relating the true response and covariates and the misclassification probabilities. For instance, an observed response of $Y_{ij} = 1$

can be obtained from: (1) true $D_{ij} = 1$ without misclassification or (2) true $D_{ij} = 0$ with misclassification. Therefore the model that describes the dependence of Y_{ij} reads as follows:

$$\rho(\boldsymbol{x}_{ij}) = Pr\left(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}, \alpha_i, \eta_i, \boldsymbol{\beta}\right) = \alpha_i \pi(\boldsymbol{x}_{ij}) + (1 - \eta_i)\left(1 - \pi(\boldsymbol{x}_{ij})\right). \quad (18.9)$$

While the main interest lies in estimating the regression parameters $\boldsymbol{\beta}$, we need to estimate also the sensitivity parameters $\alpha_1, \ldots, \alpha_7$ and the specificity parameters $\eta_1, \ldots, \eta_7$. Prior distributions are needed for all parameters. There is, however, no mathematical expression for the posterior distribution but it is possible to use a MCMC algorithm to sample from it.

The analysis and the results The effect of region, gender, age, the presence of plaque, brushing frequency, and the family smoking status on the true probability of CE in 5-year-old children was evaluated using a logistic regression model. This is done by exploring the posterior distribution of each of the regression coefficients, separately. First, the prior distributions need to be specified. For the sensitivities and specificities of the examiners, the prior can be determined from the data realized in the calibration sessions. As for the prevalence above, beta posteriors for α_i and η_i $(i = 1, \ldots, 7)$, respectively were obtained which were then used as priors in the model estimating the regression parameters. For the regression parameters we did not wish to bring in subjective information and hence used vague priors.

After specifying the priors, the likelihood, the data and the starting values in WinBugs, the MCMC process can be started. For each of the parameters a sample of the posterior is generated. Typically, a trace plot is plotted, with on the X-axis

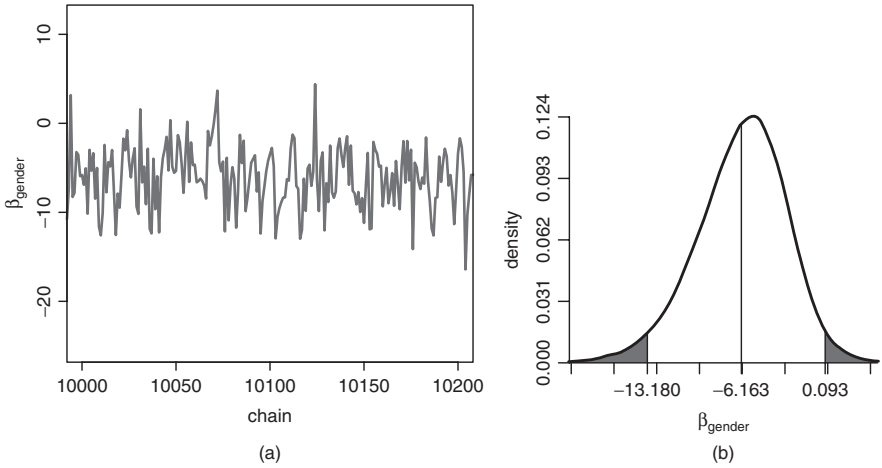


Figure 18.5 Smile for Life study: Trace plot (a) and graphical display of contour probability (b) for the regression coefficient associated to gender.

Table 18.2 Smile for Life study: Posterior mean and 95 % HPD intervals for the logistic regression coefficients with and without correction for misclassification.

Covariate	Model with correction	Model without correction
Region (B vs A)	5.459 (−3.36 ; 14.19)	−0.050 (−0.57 ; 0.47)
(C vs A)	0.875 (−9.76 ; 10.97)	0.503 (−0.01 ; 1.02)
(D vs A)	0.789 (−12.92 ; 16.27)	0.203 (−0.48 ; 0.87)
Age (years)	13.999 (2.98 ; 25.83)	1.182 (0.54 ; 1.83)
Gender (girls vs boys)	−6.163 (−13.46 ;−0.13)	−0.485 (−0.88 ;−0.09)
Presence of dental plaque		
(yes vs no)	12.670 (4.85 ; 26.62)	1.036 (0.63 ; 1.45)
Brushing frequency		
daily vs $>1\times$day	5.279 (−5.13 ; 13.38)	−0.109 (−0.63 ; 0.42)
<1 vs $>1\times$day	0.442 (−12.08 ; 9.47)	−0.198 (−0.82 ; 0.43)
Family smoking status		
not anymore vs never	−4.361 (−14.27 ; 3.85)	−0.285 (−1.00 ; 0.39)
yes/never	13.430 (3.375 ; 26.19)	1.059 (0.65 ; 1.48)

the position in the chain (the sequential number of the sample value) and on the Y-axis the sampled value. This graph gives an idea of how quickly the posterior is explored. The faster the posterior is sampled the faster one obtains an idea of the posterior from the sampled values. See Figure 18.5 for the trace plot of the regressor for gender, β_{gender}. Using WinBugs, one can also extract a posterior distribution for each of the parameters, see Figure 18.5 for (an estimate of) the posterior distribution of β_{gender}.

To test in a Bayesian context whether a regression coefficient of zero is (not) supported by the data, one could determine the appropriate Bayes' factor whereby $H_0 : \beta = 0$ is contrasted to $H_a : \beta \neq 0$. However in a complex model it is simpler to calculate the contour probabilities, see Section 18.6. The shaded area in Figure 18.5 indicates how much the null-hypothesis $\beta_{gender} = 0$ is supported by the data. The contour probabilities for the effect of the region, age, gender, presence of dental plaque, brushing frequency, and family smoking status were 0.387, 0.008, 0.045, 0.001, 0.320, and 0.004, respectively. The results indicate a non-zero effect of age, gender, presence of dental plaque, and the family smoking status on the probability of developing caries. Table 18.2 displays posterior information for the regression coefficients in the model. For the sake of comparison, the regression coefficients arising from the logistic model without correction for misclassification are also presented. Clearly, the correction for misclassification error increased considerably most of the regression coefficients in absolute value but also their uncertainty has greatly increased. When further external information on the sensitivities and specificities of the examiners (even qualitative) were available, the Bayesian approach allows to include this also in the calculations allowing to reduce the variability of the parameter estimates.

18.8.4 Emergence times of eight permanent teeth

The data and the research question The following research question emerged in the context of the ST study: 'Is the emergence time of a permanent premolar (teeth x4 and x5 in the FDI notation, with x = 1, 2, 3, and 4) affected by the caries status of the primary predecessors (teeth x4 and x5 in the FDI notation, with x = 5, 6, 7, and 8, respectively) described by a dichotomized dmft-score.'

It was also of interest to know whether the impact was the same for all premolars, which necessitates a joint (multivariate) analysis of the eight permanent premolars. Since permanent teeth emerge earlier in girls than in boys, one needs to control for gender in the analysis. Finally, Leroy *et al*. (2003) have shown that there is horizontal symmetry, i.e. the same emergence distribution can be assumed at contralateral positions (e.g., for teeth 14 and 24) and this will be used here.

The research question involves the exploration of the distribution of the emergence times. In the ST study the emergence times were left-, right- or interval-censored (see Chapter 15) and thus involves methods for the analysis of survival data. The research question was tackled for each permanent premolar separately by Komárek *et al*. (2005) and Lesaffre *et al*. (2005). For this purpose the authors used an accelerated failure time (AFT) model (see Chapter 15) with a smooth distribution estimated from the data. The AFT model was chosen to deal with the interval-censored character of many emergence times. The joint analysis of the eight premolars requires, however, a model that ties together all premolars. In this respect a Bayesian mixed effects accelerated failure time model (MEAFT) was proposed by Komárek and Lesaffre (2007) thereby introducing random intercepts (see Chapter 13). In this case the Bayesian approach made the calculations more tractable, although dedicated software had to be written.

Permanent teeth can only emerge after 5 years of age. For this reason, time to emergence $T_{i,l}$ $(i = 1, \ldots, N,\ l = 1, \ldots, n_i,\ n_i \leq 8)$ of the l-th tooth of the i-th child is modeled with 5 years of age as an onset. Hence, $T_{i,l}$ is the age when the (i, l)-th tooth emerged minus 5 years. As mentioned above, time $T_{i,l}$ is only known to lie in a time interval derived from the times when a child was examined by a dentist. To simplify calculations data augmentation was applied here, which implies that, although the true emergence time was not known, a possible time was generated at each cycle of the MCMC calculation and then further calculations were done as if the emergence time was known. This is a popular and natural procedure in Bayesian MCMC modeling. The covariates included in the model are the (vertical) position of the tooth (maxilla x4, mandible x4, maxilla x5, mandible x5), gender, CE and all two-way interactions leading to a vector $\boldsymbol{x}_{i,l}$ in expression (18.10).

The model We modelled the dependence of the (true) time to emergence $T_{i,l}$ on the covariates as

$$\log(T_{i,l}) = \boldsymbol{\beta}^T \boldsymbol{x}_{i,l} + \boldsymbol{b}_i^T \boldsymbol{z}_{i,l} + \varepsilon_{i,l} \qquad (i = 1, \ldots, N,\ l = 1, \ldots, n_i). \tag{18.10}$$

In (18.10), $\boldsymbol{\beta}$ is the vector of regression coefficients ('fixed' effects), $\varepsilon_{i,l}$ are i.i.d. error terms and $\boldsymbol{b}_i = (b_{i,1}, b_{i,2}, b_{i,3}, b_{i,4})'$ are vectors of child specific 'random' effects. More specifically $b_{i,1}, b_{i,2}, b_{i,3}, b_{i,4}$ are child specific shifts for maxilla x4, mandible x4, maxilla x5 and mandible x5. It was assumed that they follow a multivariate normal distribution with (unknown) mean and (unknown) covariance matrix. Note that the 'random' effect vector $\boldsymbol{b}_i$ which is common for all teeth of the i-th child, induces correlation between the emergence times of the teeth of a single child. Model (18.10) may seem complex, but note that if $T_{i,l}$ were fully observed (and thus not interval-censored), then this model is a mixed effects model as seen in Chapter 13. Moreover, if there were no random effects, it is just a multiple linear regression model.

To allow for some flexibility in the distribution of the emergence times, the distribution of the error term was modeled in a flexible manner, i.e. as a normal mixture. This approach is described in detail in Komárek and Lesaffre (2007).

Our model allows the effect of CE on the emergence to differ between boys and girls and between maxillary and mandibular premolars. The effect itself is quantified by an acceleration factor (AF) which is given as an exponential of a proper combination of β parameters. Note that the closer AF is to one for a particular covariate combination, the less impact this covariate combination has on the emergence process. Acceleration factors evaluating the effect of a decayed primary premolar are given in Table 18.3 and determine by which factor the emergence, as measured from 5 years of age, is accelerated ($AF < 1$) or decelerated ($AF > 1$) for the tooth with caries experience on the primary predecessor compared to the tooth on the same position with sound predecessor.

The analysis and the results Posterior summary measures for acceleration factors are given in Table 18.3, together with the contour probabilities. It is seen that when the primary predecessor is decayed the emergence of the permanent successor is accelerated in the case of maxillary teeth. For example, the emergence of maxillary permanent first premolars in boys is accelerated by a factor of 0.9553 with 95 % posterior uncertainty given by (0.9385, 0.9716) in case of a primary premolar with CE. For the mandibular teeth, a slight effect (AFs are close to 1) is observed but only for the first premolar on boys.

Further, Figure 18.6 shows the estimated error distribution (of $\varepsilon_{i,l}$). Together with the normal distribution of the random effects, the estimated emergence curves (1-survival curves) for all permanent premolars can be visualized and thereby the effect of the covariates quantified. Figure 18.6 shows these emergence curves for the maxillary first premolar, separately for boys and girls and for the two CE groups.

The motivation of using the Bayesian approach was primarily computational and especially because MCMC techniques together with data augmentation render the modeling feasible. Indeed, integral calculations would imply a formidable task.

Table 18.3 Signal-Tandmobiel® study: posterior median, 95 % HPD interval and associated contour probabilities for acceleration factors for the two genders and contralateral teeth evaluating the effect of a decayed primary premolar. Based on data from Komárek and Lesaffre (2007), published by *Statistica Sinica*.

Maxilla 4		Maxilla 5	
Girl	Boy	Girl	Boy
0.9655	0.9553	0.9790	0.9688
(0.9491, 0.9816)	(0.9385, 0.9716)	(0.9618, 0.9965)	(0.9509, 0.9863)
$P < 0.001$	$P < 0.001$	$P = 0.019$	$P < 0.001$
Mandible 4		**Mandible 5**	
Girl	Boy	Girl	Boy
0.9903	0.9801	1.0015	0.9910
(0.9734, 1.0067)	(0.9632, 0.9970)	(0.9840, 1.0195)	(0.9722, 1.0099)
$P = 0.255$	$P = 0.021$	$P = 0.870$	$P = 0.353$

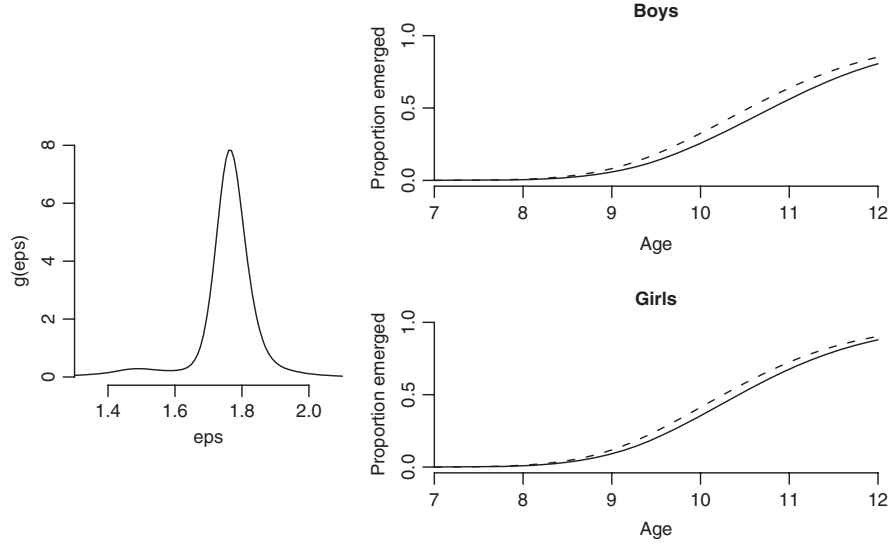


Figure 18.6 Signal-Tandmobiel® study. Estimated error distribution and cumulative distribution functions (emergence curves) for maxillary first premolars (solid line for the premolars with a sound primary tooth, dashed line for the premolars with a primary tooth showing CE). Based on data from Komárek and Lesaffre (2007), published by Statistica Sinica.

18.9 Conclusions

The Bayesian approach differs considerably in philosophy as well as in computations from the classical statistical approach. It is claimed that the Bayesian paradigm is more natural than the classical frequentist since it uses only the observed data at hand to draw scientific conclusions. Further, the Bayesian approach allows all relevant information to be incorporated into the statistical analysis. Finally, in the last two decades the computational capabilities have been greatly extended through the introduction of MCMC techniques which now allow the tackling of complex modeling problems. Unfortunately, the Bayesian methodology is highly technical and definitely needs statistical support. There are tools available that allow the OH researcher to get acquainted with the Bayesian terminology such as the FirstBayes program (http://www.tonyohagan.co.uk/1b/). The literature on Bayesian methods is also mainly technical, but a good reference for the medical/dental reader is e.g. Spiegelhalter *et al*. (2004). The more ambitious reader could consult the Bayesian literature, e.g. Gelman *et al*. (2004) or the book in preparation of the first author (Lesaffre and Lawson, 2009) which contains a number of dental examples.

Acknowledgements

The first author acknowledges the partial support of the Research Grant OT/05/60, Catholic University Leuven. Arnošt Komárek's work was supported by the MSM 0021620839 grant, Ministry of Education, Youth and Sports of the Czech Republic. Alejandro Jara's work was supported by the Fondecyt grant 3095003.

References

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* **53**, 370–418.

Box G & Tiao G (1973) *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley. Reprinted by John Wiley & Sons, Inc., New York, in 1992 in the Wiley Classics Library Edition.

Declerck D, Leroy E, Martens L, *et al*. (2008) Factors associated with prevalence and severity of caries experience in pre-school children. *Community Dentistry and Oral Epidemiology*, **36**(2), 168–178.

Fisher R (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher R (1935) *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Gelfand A & Smith A (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman A., Carlin JB, Stern HS and Rubin DB (2004) Bayesian Data Analysis. Boca Rotan: Chapman & Hall/CRC.

Geman S & Geman D (1984) Stochastic relaxation, Gibbs disributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–41.

Goodman S (1993) P values, hypothesis tests, and likelihood approaches for epidemiology of a neglected historical debate. *American Journal of Epidemiology* **137**, 485–500.

Goodman S (1999a) Toward evidence-based medical statistics. 1:The P value fallacy. *Annals of Internal Medicine* **130**, 995–1004.

Goodman S (1999b) Toward evidence-based medical statistics. 2:The Bayes factor. *Annals of Internal Medicine* **130**, 1005–13.

Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hubbard R & Bayarri M (2003) Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician* **57**, 171–82.

Komárek A & Lesaffre E (2007) Bayesian accelerated failure time model for correlated censored data with a normal mixture as an error distribution. *Statistica Sinica* **17**, 549–69.

Komárek A, Lesaffre E & Hilton JF 2005 Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics* **14**, 726–45.

Leroy R, Bogaerts K, Lesaffre E & Declerck D (2003) The emergence of permanent teeth in Flemish children. *Community Dent Oral Epidemiology* **31**, 30–9.

Lesaffre E & Lawson A (2009) *Bayesian Methods in Biostatistics*. John Wiley & Sons, Inc. New York (in preparation).

Lesaffre E, Komárek A & Declerck D (2005) An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research* **14**(6), 539–52.

Lilford R & Braunholtz D (1996) The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal* **313**, 603–7.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A & Teller E (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–91.

Neyman J & Pearson E (1933) On the problem of the most efficient test of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A* **231**, 289–337.

O'Hagan A, Buck C, Daneshkrah A, *et al*. (2007) *Uncertain Judgments: Eliciting Expert's Probabilities*. John Wiley & Sons, Inc. New York.

Spiegelhalter D, Abrams K & Myles J (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, Inc., New York.

Tan SB, Chung YF, Tai BC, Cheung YB & Machin D (2003) Bayesian approaches to randomised trials. *Controlled Clinical Trials* **24**, 110–21.

Part V

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

19

Examples from oral health epidemiology: the Signal Tandmobiel® and Smile for Life studies

Dominique Declerck, Emmanuel Lesaffre, Roos Leroy and Jackie Vanobbergen

19.1 Introduction

The study of the distribution of oral diseases, related risk factors, and mechanisms to control oral health problems is essential for the planning, organization and evaluation of oral health promotion activities and oral health services delivery. In the field of oral health, complex interactions and associations may exist between e.g. social or psychological factors and the level of oral hygiene and dietary habits. This necessitates the development and application of multifactorial and complex disease models.

In this chapter, we describe two large epidemiological surveys in children regarding oral health promotion and oral health. The main focus of this contribution is to explain the study set-up and conduct of both surveys in detail and to present some of the analyses that were performed based on these datasets. For additional details, we refer to published reports (see below).

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

19.2 The Signal Tandmobiel® study

19.2.1 Introduction

Since 1980, a number of oral health surveys have been undertaken in Flanders, the northern part of Belgium (for an overview, see Vanobbergen *et al*., 2001a). The majority of these surveys were cross-sectional, limited to a specific geographical area or target group and confined to a specific age group. Most surveys included only low numbers of subjects in poorly defined samples.

In order to obtain more reliable information on the oral health of Flemish primary schoolchildren, the Signal Tandmobiel® project was launched in Flanders in 1996. The aims of this longitudinal study were: (1) to assess the oral health condition and its determinants in primary schoolchildren between the ages of 7 and 12 years; and (2) to implement and evaluate an oral health promotion programme offered to these children.

19.2.2 Study design and conduct

19.2.2.1 General set-up

In order to collect longitudinal data on the oral health condition of primary schoolchildren, a cohort of Flemish children was examined annually for a period of 6 years (between 1996 and 2001) (see also Table 19.1). At the start of the project, the children were about 7 years old.

Table 19.1 General set-up of the Signal Tandmobiel® project: samples and examinations in different survey years.

Survey Year	1996	1997	1998	1999	2000	2001
Group A						
Longitudinal follow-up, oral health promotion programme	X	X	X	X	X	X
Group B						
Longitudinal follow-up, NO oral health promotion intervention	X					X
Group C						
Cross-sectional, separate samples		X				
			X			
				X		
					X	
						X

X = clinical exam + questionnaire

The cohort consisted of two groups: a group being examined yearly and receiving oral health education at each of these occasions (A-sample) and a (smaller) group for which an exam took place only at baseline (age 7 years) and at the end of the study (age 12 years) (B-sample). The latter group received no oral health education and served as a control for the evaluation of the impact of the intervention. In addition, in each survey year a control group of age-matched children was examined (C-samples). These children were examined only once and served as controls for cross-sectional comparisons at different ages. For further details on the set-up of the project, we refer to Vanobbergen *et al*. (2000).

19.2.2.2 Data collection: sampling procedure

Samples were randomly selected through cluster (i.e. school) sampling without replacement, stratified by province and type of educational system (state funded, municipal and private institutions). Schools were selected with a probability proportional to their size; i.e. the number of children in the first year of primary school. The A-sample consisted of 4468 children (2315 boys and 2153 girls) from 179 different primary schools, representing 7.3 % of Flemish children born in 1989. The mean age of the children at the first examination was 7.08 (SD = 0.41) years. Every survey year, between 452 and 1112 children were unavailable for examination. Major reasons for nonparticipation were checked and were in almost all cases unlikely to be related to the objectives of the study (i.e. illness or absence on the day of the examination or change of school). The B-sample consisted of 820 pupils, from which 676 children were examined in the final survey year (82.4 %); C-samples comprised approximately 600 children each.

19.2.2.3 Questionnaires

Information on reported oral health, dietary habits and socioeconomic background of the children was obtained from the parents of the child and from the school health care centre, using structured questionnaires.

When data are collected by interviews or questionnaires, *limitation in recall* should not be overlooked. For example, the parents of the examined children were questioned about the age tooth brushing started and the use of systemic fluoride supplements at the time their son or daughter was already seven years old. Persson and Carlgren (1984) evaluated dietary assessment techniques in infancy and childhood and observed that reliability concerning the reported end of breastfeeding decreased over time. They suggest continuous monitoring or repeated interviews to obtain valid and reliable feeding data. Within this project, the investigators did not have the possibility to assess the reliability of the information provided by the parents. It is important to consider this issue when interpreting findings. Moreover, some parents may have overestimated the use of systemic fluorides and oral hygiene habits out of *motives of social desirability*. Thus, some children may have been categorized in the wrong group. Misclassification of ‘exposure’ usually results in differences between groups being veiled to some extent. Hence, the true risk or benefit from exposure may be underestimated.

The introductory paragraph of the questionnaire was used to inform the parents about the project and to obtain parental consent for data collection. The following items were included in the questionnaire: oral hygiene habits, use of fluoride supplements, dental attendance, dietary habits, history of traumatic injuries to the teeth, ethnic origin, socio-economic background (based on the occupational level of the parents), medical background of the child (if relevant for oral health). Questionnaires were distributed at school by the teachers.

19.2.2.4 Clinical examinations

Examinations were carried out by one of 16 dentist-examiners specifically trained for this purpose. The selection of examiners, all general practitioners with at least five years of clinical experience, was based on the results of an intensive three day training programme that included theoretical sessions, slide demonstrations and practical exercises.

In order to maintain an adequate level of reliability of the recordings and a sufficient level of agreement amongst examiners, calibration sessions were organized annually. (Weighted) kappa values for the scoring of caries experience ranged between 0.72 and 0.91 and were higher than 0.80 for 13 out of 16 dental examiners. For further discussion on the use of kappa values, we refer to Section 19.2.3.3 or Lesaffre *et al*. (2004). The children were examined while seated in a standard dental chair that was installed in a fully equipped van parked on school premises. The examination was carried out using a disposable mouth mirror and WHO/CPITN type E probe (Prima Instruments, Prima Dental Group, Gloucester, UK). Because the examination was not connected to routine dental check-ups, no radiographs were taken.

In order to assess the level of oral hygiene, the presence of dental plaque was determined. For this purpose, two different indices were used. Buccal surfaces of teeth 16, 21, 24, 36, 44 and 41 (or primary alternatives 55, 61, 64, 75, 84 and 81) were scored according to the index described by Silness and Loë (1964). When present, permanent teeth were examined. Because this index does not include plaque accumulation on occlusal surfaces, the most caries prone surfaces, a separate registration was done on these surfaces in the permanent first molars using a simplified version of the index described by Carvalho *et al*. (1989). The gingival condition was measured using the Sulcus Bleeding index method described by Mühlemann and Son (1971). The buccal surfaces of the same index teeth as described above were used.

Teeth were examined for presence of caries experience after removal of debris and plaque using cotton gauze or a dental probe. Compressed air was used to assure good visibility. The condition of the teeth was recorded at tooth surface level using the guidelines proposed by the British Association for the Study of Community Dentistry (BASCD) (Pitts *et al*., 1997). Decay was recorded at the level of cavitation. In addition to caries experience, developmental defects of enamel and clinical signs of fluorosis were registered. For additional detail, we refer to Vanobbergen *et al*. (2000).

19.2.2.5 Data management and analysis

Clinical, as well as questionnaire data, were entered on-site into an electronic database using Dental Survey Plus version 4.50B (Providence software Services, Bristol, England). Data entry was the responsibility of the person assisting the dentist-examiners. Clinical and questionnaire data were converted into SAS® data files (SAS Institute, Cary, NC, USA)(version 8.2).

19.2.3 Examples of analyses based on the Signal Tandmobiel® database

Below we present three examples of analyses based on the data obtained in this oral health survey. In Chapter 18, additional applications are presented.

19.2.3.1 Example 1: Evaluation of a caries risk assessment model

Based on the data obtained in the first survey year, the impact of a set of risk indicators on caries prevalence in the primary dentition of 7-year old Flemish children was defined (Vanobbergen *et al*., 2001b). The validity of the risk indicators identified in the cross-sectional set-up was then evaluated for the prediction of caries development in the first permanent molars ($DMFS_6$-scores, caries experience scores limited to the four first permanent molars) by age 10.

Baseline data were available for 3303 children, born in 1989, of whom 2691 (81 %) were again examined at the fourth examination (age 10 years).

Different outcome measures for assessing caries development were calculated, using data collected for the four first permanent molars: net caries increment, cumulative caries incidence and caries incidence density. Net caries increment was calculated for each child by subtracting the baseline $DMFS_6$ score from the last available $DMFS_6$ score. Cumulative incidence represented the proportion of children who developed new caries lesions within the observation period. This proportion was calculated by counting the number of children whose caries increment exceeded zero during the observation period, and then dividing by the total number of children. Caries incidence density represented the average change in caries experience status per unit of time relative to the number of surfaces at risk, based on the annually recorded emergence stages of the teeth (range 0 to 4, from non erupted to fully erupted and full occlusal contact). Incidence density was expressed in newly affected surfaces (increment) per 100 surface years (Slade & Caplan, 2000).

As prediction variables, baseline data on oral health status, oral hygiene level, reported oral health behaviour and socio-demographic factors were used. Based on this model, the probability of being affected by caries in the primary dentition by the age of 7 was determined for each child and used as an indicator for future caries development on first permanent molars.

In order to quantify caries development in first permanent molars between the ages of 7 and 10, the cumulative incidence was calculated for the whole sample and for the three classes of children based upon the quartiles obtained using the

above-mentioned prediction variables (<Q1; Q1–Q3;>Q3). A stepwise logistic regression analysis was performed with net caries increment as the outcome measure, dichotomized as no or 1 additional surface affected versus 2 or more, and adjusted for the real time at risk using emergence data. To get an idea of the predictive power, different dichotomies for both predictor and outcome were formed and summarized as ROC-curves (receiver operating characteristic curves) and an index called 'area under the curve' (AUC), see Chapter 12.

Results showed that baseline dmfs and occlusal and buccal plaque indices were highly significant for having a positive caries increment in permanent first molars with respective odds ratio's of 1.07 (95 % CI: 1.05–1.08), 1.43 (95 % CI:1.28–1.80) and 1.35 (95 % CI: 1.08–1.68). Brushing less than once a day and the daily use of sugar-containing drinks between meals were confirmed as risk factors (OR 2.43 (95 % CI: 1.70–2.96) and 1.25 (95 % CI: 1.00–1.56), respectively). Logistic regression analysis using the different dichotomies provided a ROC-curve resulting in an AUC (area under the curve) of 0.72 which indicates that the risk marker showed only moderate predictive power. None of the sociodemographic and behavioural variables had enough predictive power at the group level to be useful for identifying caries susceptible children. Even the predictive power of dmfs at baseline was considered modest. For further details, we refer to Vanobbergen *et al*. (2001c).

19.2.3.2 Example 2: Timing and sequence of emergence of permanent teeth

At each annual examination of the Signal Tandmobiel® project, all permanent teeth were scored with respect to their emergence status. A tooth was scored as emerged when at least one cusp was visible in the mouth. Teeth extracted for orthodontic reasons were recorded as having emerged.

A logistic survival analysis on the logarithmic transformation of age was performed to calculate median age at emergence and 95 % confidence intervals for all permanent teeth (Leroy *et al*., 2003a). This model assumes that log(age) has a logistic distribution (somewhat wider tails than normal distribution) and proved to fit best the emergence times among many survival models that were evaluated, see Leroy *et al*. (2003a). In order to evaluate the statistical significance of the differences in the median emergence ages of contralateral and opposing teeth, the GEE approach (see Chapter 13) was extended to model multivariate interval censored emergence times (Bogaerts *et al*., 2002). Multivariate survival analyses with pairwise tests were performed, taking into account the dependence of teeth within an individual, as was treated in Chapter 15. General trends, also reported elsewhere, were observed: significantly earlier emergence in girls than in boys, most mandibular teeth emerge before their maxillary antagonists and the differences in median emergence ages between contralateral teeth are clinically negligible. Figure 19.1 illustrates the emergence curves of the right maxillary and mandibular central incisors, canines and first premolars for boys and girls separately.

In order to evaluate the effect of caries experience in a primary molar (decayed and/or restored versus extracted) on the timing of emergence of its permanent

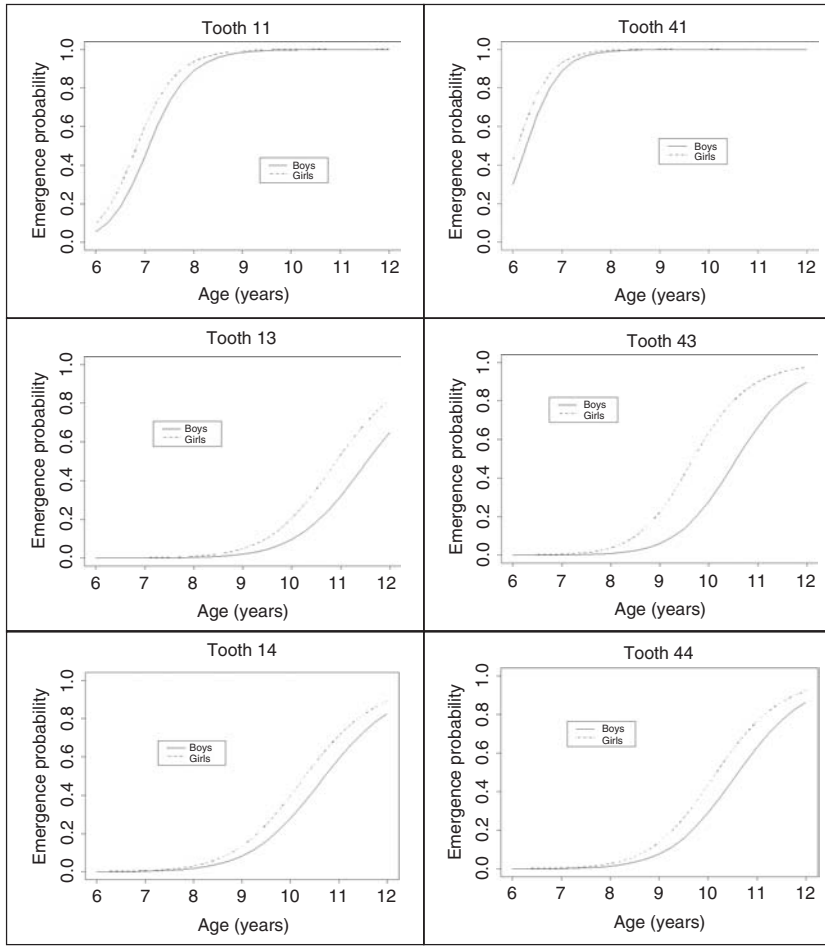


Figure 19.1 Emergence curves for the right maxillary and mandibular central incisors, canines and first premolars. Leroy R, Bogaerts K, Lesaffre E, Declerck D. The emergence of permanent teeth in Flemish children. Community Dent Oral Epidemiol 2003; ***31****, 30–9.*

successor, caries experience in the primary molars was added to the model as a covariate. Once a primary molar had been scored as decayed, restored or extracted, it was assigned to the category of caries-positive teeth.

In a first analysis, the emergence of premolars with a predecessor without caries experience (dmft = 0) was compared with the timing of emergence of premolars with a predecessor with caries experience (dmft > 0). In a second analysis, the caries-positive group was further subdivided into teeth extracted due to caries experience (mt > 0) versus decayed and or restored, but never scored as extracted because of caries experience (dft > 0). Per tooth type, 12 statistical tests (i.e. for

each of the four teeth: dmft = 0 versus dmft > 0; dmft = 0 versus dft > 0 and dmft = 0 versus mt > 0) were performed. A Bonferroni correction for multiple testing (see Chapter 10) was applied. All analyses were performed for boys and girls separately.

Results indicate that caries experience in a primary molar, irrespective of whether the tooth was extracted or not, resulted in a statistically significant acceleration of the emergence of the successor. The effect was more pronounced in boys (4–8 months), than in girls (2–4 months). When a primary molar was decayed and/or restored but was never scored as extracted due to caries experience (i.e dft > 0), the emergence of its successor was also significantly accelerated (3–8 months). Premature loss of maxillary primary molars (i.e. mt > 0) resulted in a significant acceleration of the emergence of the premolars (10–19 months). In the mandible, the effect of (premature) extraction was negligible. For more details, we refer to Leroy *et al*. (2003b).

Based on mean and median emergence ages, 'classical' tooth emergence sequences have been suggested and presented in text books on paediatric dentistry and orthodontics as 'desirable emergence sequence for permanent teeth'. However, in the clinical paedodontic and orthodontic settings one is much more interested in individual sequences of tooth emergence, as individual emergence sequences can not be evaluated nor predicted based on mean emergence orders. In addition, presenting emergence sequences based on mean or median emergence ages may be misleading as insight into the variability of the phenomenon is limited.

Even though the study of the variability in tooth emergence order is quite desirable, its establishment is not straightforward. In order to obtain information on individual sequences in a population, prospective data from a large sample are needed. The statistical analysis is quite complicated. For example, the emergence patterns in one quadrant involve a seven-dimensional model that takes into account the interval-censored character of the emergence times. Further, to examine the dependence of the emergence structure on covariates, the covariance structure should be modeled as a function of covariates. Furthermore, for the 7 permanent teeth in a quadrant (excluding the wisdom tooth), there are $7\times6\times5\times4\times3\times2\times1 = 5040$ different possible sequences of emergence. If the exact emergence time of each tooth were recorded, then for each child the sequence of emergence would be known and the prevalence of a sequence easy to establish. A common approach to handling interval-censored data is to assign a particular value to the event of interest (e.g. midpoint of the time interval in which the event took place) and then proceed as if the data were known exactly. However, this can lead to biased and misleading results, as reported in Chapter 15.

For each of the 5040 possible sequences the probability (prevalence) that teeth emerge in that sequence was estimated, based on a model for the unobserved latent (true) emergence times. Specifically, it was assumed that the true emergence times of the seven teeth follow a multivariate normal distribution. Upon the estimation of the mean and covariance structure of the normal distribution of the emergence times, the prevalence for each sequence of the 5040 different sequences was established. Parameter estimation was done using the Bayesian approach which deals

elegantly with interval-censored data (see Chapter 18). The technical details on how the model was implemented were previously published (Cecere *et al*., 2006).

The analyses indicated that no sequence in a quadrant was common to more than 19 % of the sample (Leroy *et al*., 2008a). When only those emergence sequences with a prevalence of 1 % or more are considered, 21 variations can be expected in the maxilla and 15–22 variations in the mandible. These variations 'cover' 84–88 % of all sequences. The presentation of emergence sequences solely based on means or medians should thus be regarded as misleading. It was also observed that caries experience on the deciduous teeth distorts the emergence process of the permanent teeth.

19.2.3.3 Example 3: Analysis of caries experience taking into account inter-observer bias and variability

In oral health surveys, clinical measurements are often collected using multiple examiners. This raises a concern about inter-examiner variability. The problem is usually dealt with by simply reporting kappa values (Cohen, 1960). High values of κ indicate high agreement levels between examiners' scoring. However, this statistic has important limitations. In addition, when a gold standard or benchmark scorer is available, sensitivity and specificity calculations of each examiner vis-à-vis this gold standard (or benchmark scorer) are preferred.

We now demonstrate the limitations of reporting the κ-statistic when analyzing oral health screening data collected in the Signal Tandmobiel® study.

Data from three different calibration exercises were available from at least 12 children who were examined by the 16 dental examiners and the benchmark scorer. Based on this validation dataset, kappa-values (weighted) were calculated. Values ranged between 0.72 and 0.91 and were higher than 0.80 for 13 out of 16 of the dental examiners.

Earlier work showed a clear East–West gradient in the level of caries experience among 7-year olds in Flanders (Figure 19.2). This gradient was confirmed by a highly significant regression coefficient for the x-coordinate of the geographical location of the school based on a (Bayesian) ordinal logistic regression analysis (for more details, see Lesaffre *et al* 2004). However, a similar gradient in the scoring behaviour of the 16 dental examiners could be seen in the validation dataset (Figure 19.2). Therefore, it was questioned whether the observed geographical gradient in caries experience scores was genuine or caused by different scoring behaviours of the examiners. The fact that acceptable κ-values were obtained is not helpful in answering this question.

Therefore the under- and overscoring behaviours of the dental examiners vis-à-vis the benchmark examiner, were incorporated into the ordinal logistic regression model using an approach that corrects for misclassification, see e.g. Chapter 16. When comparing models with and without this correction, it was shown that the East–West gradient was again highly significant (in a Bayesian sense), but was estimated with less precision. Further analyses, including other covariates like regional deprivation indices, local fluoride level of the drinking

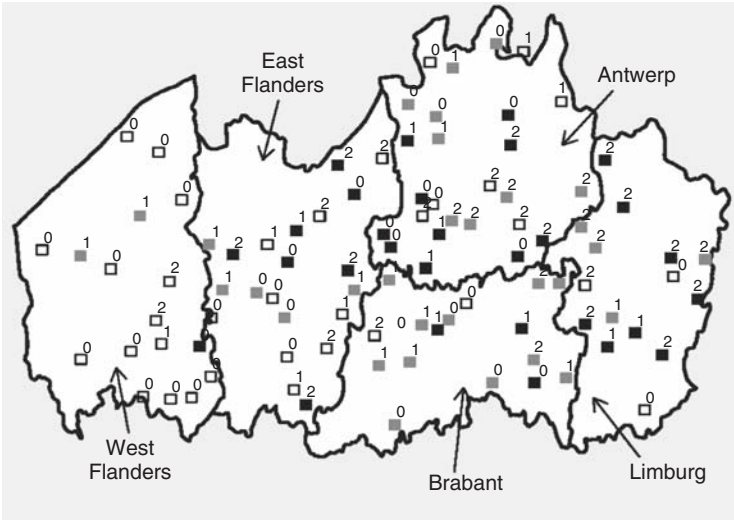


Figure 19.2 The Signal Tandmobiel® study: Map of Flanders with level of caries experience and over- and under-scoring of dental examiners. Caries experience was split into 3 categories according to quartiles of the mean dmft scores obtained per school and coded as 0 (minimum to Q1), 1 (Q1 to Q3) or 2 (above Q3). The over- or underscoring of the examiner is indicated with the symbols □, ■ and ■. The symbol □ signifies that the dental examiner scoring the respective school underscored 5 % to 15 % compared to the benchmark examiner in the calibration exercises. The symbol ■ signifies between 5 % under- and 5 % over-scoring, and the symbol ■ signifies at least 5 % over-scoring (up to 18 %).

water and others, did not remove the geographical gradient. In this way, it could be ruled out that the observed geographic East–West gradient in caries experience of 7-year olds in Flanders was due to bias induced by variability in the scoring of the different examiners. For more details, we refer to Lesaffre *et al.* (2004).

19.3 The Smile for Life study

19.3.1 Introduction

Epidemiological data on the oral health of pre-school children are limited. Available information is usually from 5-year olds; younger age groups are more difficult to reach (not yet attending kindergarten), and they present a challenge to examine because of variable levels of cooperation.

On the other hand, it has been shown in the Signal Tandmobiel® study and by others that the condition of the primary dentition is highly predictive for the condition of the permanent dentition (Hausen, 1997; Vanobbergen *et al.*, 2001c;

Leroy *et al*., 2005). Therefore, profound understanding of mechanisms of oral disease development at a very young age is of utmost importance. Based on this knowledge, targeted oral health promotion interventions can be developed.

In order to determine the oral health situation of young children and to get more insight in the factors associated with disease prevalence and distribution, the *Smile for Life (Tandje de Voorste)* study was launched in 2003. Based on the information obtained from the screening of 3- and 5-year olds at baseline (oral examination and questionnaire completed by the parents), an oral health promotion intervention programme was developed and implemented in 1000 newborn children and their parents. Evaluation of the impact of the intervention on oral health and related behaviours of this cohort is performed when the children reach the age of 3 and 5. Another 1000 children, not residing in the intervention area, are examined at the same ages and serve as controls. This project is the result of a multidisciplinary effort involving input from the field of oral health sciences, youth health care, health psychology and biostatistics.

The objective of this study was the evaluation of a specifically designed oral health promotion intervention programme with very young children and their parents.

19.3.2 Study design and conduct

19.3.2.1 General set-up

The *Smile for Life* project can be subdivided into two parts: (1) baseline data collection and development of an oral health promotion intervention and (2) implementation and evaluation of the intervention. The aim of baseline data collection was to examine the prevalence and severity of caries experience in Flemish preschool children and to assess the association of disease distribution with oral hygiene levels, reported oral health behaviours and sociodemographic factors. Based on the information collected, an oral health promotion intervention was developed.

In the second part of the project, the intervention is implemented and evaluated. For this purpose, a cohort of children is followed from birth until the age of 5 years. The impact of the intervention will be assessed using a comparison of selected oral health outcome measures and related behaviours between intervention and control groups.

19.3.2.2 Baseline data collection: sampling

Baseline data collection was undertaken in four distinct geographical areas in Flanders, coinciding with regional activity centres of *'Child & Family'*. This organization, subsidized by the Flemish government, offers medical and educational support to all young parents and their children between birth and the age of 2.5 years. Services (including vaccinations) are free of charge. Over 95 % of parents make use of this offer, irrespective of socio-economic background.

In each of these four regions, approximately 30 % of 3- and 5-year olds were included in the sample, yielding a total sample of 1250 3-year olds (born in 2000)

and 1283 5-year olds (born in 1998). Sampling was performed based on listings of kindergartens in the areas considered. The selection of individual children was not possible because of ethical, practical and economical reasons. In Flanders, attendance of kindergarten is high with 99 % of 3-year olds and 98 % of 5-year olds attending (data for school year 2002–03, source: Ministry of the Flemish community, Education Department).

Kindergartens were selected using stratified cluster sampling without replacement. The target population was divided into three strata, representing the different types of educational system (state funded, municipal and private institutions) and considering an equal spread in rural and urban regions. Whenever a school was selected, all 3- and 5-year-old children attending were included. Since the probability of selection of kindergartens was proportional to their size, each child had the same probability of being selected. Five out of 80 selected schools refused participation.

19.3.2.3 Oral health promotion intervention

The four *Child & Family* regions where baseline data were collected were included in the second part of the study. Two regions were designated as intervention regions and the two remaining ones served as control regions. The control regions were selected using descriptive information (number of births, socioeconomic situation, Child & Family attendance, ...) and their geographical location (at least 100 km away from intervention areas in order to limit transfer of information and materials between regions).

All newborn children (and their parents) from both intervention regions were offered the oral health promotion programme. In the control regions, no extra information regarding oral health was provided by *Child & Family*, except the information already included in their standard offer.

In each of the regions, the first 500 children born after 1 October 2003 were included in the project (approximately 50 % of annual births in the respective regions). Nurses of *Child & Family* presented the parents of these newborns with a questionnaire regarding oral health and related items. Parents had the possibility to refuse participation. Newborns with medical problems that could affect their oral health and children whose parents did not have enough understanding of the language in order to be able to complete the questionnaire were excluded. In case of twins, the child whose first name ranked first in alphabetical order was included. The oral health promotion intervention programme was developed based upon a combination of literature findings and the results obtained from the analysis of data collected at baseline (see Declerck *et al*., 2008). Once the content of the intervention was determined, the different messages were integrated into the already existing care programme offered by Child & Family to parents and their young children. The nurses played an important role in this process, as their experience in working with young parents was considered essential in order to transfer the oral health messages in an efficient way. At each contact with the nurse or physician of *Child & Family* (4 visits at home and 11 visits at the consultation office over a period

of 2.5 years), oral health related items were presented to the parents. The message was supported by offering specifically developed educational tools (flyer with oral health information for pregnant women, child booklet, rinsing cup, toothbrush and toothpaste and place mat). This programme was only offered to the children in the intervention regions. Children from both the project and control regions were examined at the age of 3 in 2007 and are examined again at the age of 5 in 2009, in order to assess the impact of the intervention.

19.3.2.4 Questionnaires

In this project, several questionnaires were used.

At baseline data collection, not only were the parents asked to complete a questionnaire, but also the personnel of *Child & Family* and different groups of (oral) health care workers (general physicians, paediatricians, gynaecologists, nurses, pharmacists, teachers, dentists, ...) in the four respective regions. Questionnaires for the parents of 3- and 5-year olds were distributed at schools by the teachers and were accompanied by a letter explaining the purpose of the survey. Parents were asked to complete the questionnaire and return it to the teachers. The questionnaire included the request for permission from the parents to have their child examined by a dentist at school. Completed questionnaires were returned by more than 89 % of the parents of children in both age groups.

Questionnaires to be completed by health care workers were delivered by postal mail, based on listings of registered health care workers in the respective regions. A pre-stamped return envelope was included. The overall response rate for data collection at baseline was 49 %.

In the second part, the intervention part of the project, parents were asked to complete a questionnaire shortly after their child was born (within the first 4–8 weeks). The nurses from *Child & Family* were responsible for the distribution and retrieval of these documents. About 99 % of parents in the intervention group and 91 % of parents in the control group completed the questionnaire. In addition, questionnaire information was collected during the follow-up period when the child was 3 and 5 years. Since the oral health examinations were carried out at school, the distribution of the questionnaires was organized in the same way as was the baseline data collection. At the age of 3 years, 62 % of the parents returned the questionnaire. Here also, the request for permission from the parents to examine the oral cavity of their child was integrated in the questionnaire.

Questionnaires were also sent to different groups of oral health care workers, using the same approach as described above (baseline).

All questionnaires consisted of structured questions (pre-set categorical responses), with only a few open-ended questions. They included the following items: oral hygiene and dietary habits; other oral health related behaviours; sociodemographic information and determinants of oral health related behaviours. Questionnaires were validated and tested in a pilot study. For additional information on this procedure, we refer to Defranc *et al*. (2008). For more detail on the content of the questionnaires, see Declerck *et al*. (2008).

19.3.2.5 Clinical assessment

All children were examined at the school premises. In order to minimize the possibility of extra brushing efforts, parents were not informed in advance of the exact date of the oral health examination. At baseline, the children were examined by one of 8 dentist-examiners. These dentists were selected from a group of dentists who had previously participated in oral epidemiological surveys and/or volunteered for participation in this project. All dentists were trained in the examination methodology. The overall set-up of the project was explained, organizational aspects were agreed upon and the various clinical variables to be assessed were described and illustrated using slides. At the end of the training period, a calibration exercise was organized using clinical pictures. Differences in scoring behaviour were discussed. In addition, a calibration was organized consisting of the examination of a group of children of the same age but not participating in the *Smile for Life* project. Sensitivity and specificity in the scoring of caries experience was estimated for each dental examiner versus the benchmark examiner (DD). Sensitivity scores ranged between 0.57 and 0.71, specificity scores ranged between 0.87 and 1.00.

The children were examined seated on an ordinary chair and using a mouth mirror and probe. For this purpose, a mouth mirror with built-in light source was used (Mirrorlite™ by Defend® from Medident, Belgium) and a WHO/CPITN type E probe (Prima Instruments, Prima Dental Group, Gloucester, UK). Oral hygiene level was assessed using the method described by Alaluusua *et al*. (1994). This index records the presence/absence of dental plaque accumulation based on a visual examination. Recordings were made on the buccal surfaces of teeth 52, 55, 72 and 75 (as proposed by Carvalho *et al*., 1998). A global plaque score was obtained for each child by calculating the mean score of the surfaces examined.

Caries experience was recorded using the criteria proposed by the British Association for the Study of Community Dentistry (BASCD) (Pitts *et al*., 1997). When necessary, teeth were dried and/or cleaned using cotton rolls. Children were not asked to clean their teeth immediately before the examination. No radiographs were taken. Carious lesions were assessed in such a way that reporting and analysing of data was possible at the level of the initial lesion (demineralization level, d1 level) and at the cavitation level (d3 level). Tooth surface was used as the unit of observation. Caries experience was summarized using the dmft/dmfs variable as described by Klein & Palmer (1938).

Other variables included in the clinical assessment of the children are: gingival inflammation, presence of mucosal lesions, pellicle discoloration, developmental defects of tooth structure, erosion, damage due to trauma and malocclusions. Details on these variables are not described in this chapter.

19.3.2.6 Data management and analysis

Questionnaire data were entered into a database using Excel (Microsoft). All baseline data were entered twice, by two different persons. Excel Compare™ was used to check agreement between both databases. In case inconsistencies were detected,

the original questionnaires were checked and the necessary corrections were made so that two identical files were obtained.

Clinical data were entered on-site into an electronic database using Dental Survey Plus version 4.50B (Providence software Services, Bristol, England). Data entry was the responsibility of the nurse assisting the dentist-examiners.

Clinical and questionnaire data were first converted, then merged into SAS® data files (SAS Institute, Cary, NC, USA) (version 8.2).

19.3.3 Example of analysis based on the *Smile for Life* database: parental smoking behaviour and caries experience

The aim of this analysis was to investigate the association between parental smoking behaviour and caries experience in their young children. Possible confounding effects from socioeconomic status or oral health-related behaviour were considered.

Chi-squared tests were performed to evaluate the association between family smoking status and other covariates. Simple and multiple logistic regression analyses were performed for all covariates with caries experience (expressed as a binary outcome, i.e. $d_3mft = 0$ versus $d_3mft > 0$) as the outcome variable. Analyses were corrected for examiner misclassification, based on the results of the calibration exercises. P-values at or below 0.05 were considered statistically significant.

In both age groups, a smoking habit was reported by 30 % of the parents. Simple logistic regression analysis with caries prevalence as the response revealed that parental smoking was a significant covariate. After controlling for age, gender, socio-demographic characteristics, oral hygiene and dietary habits, the effect of family smoking status was no longer significant in 3-year old children (OR = 1.98; 95 % CI: 0.68–5.76). In 5-year olds the significant relationship between parental smoking behaviour and caries experience persisted after adjusting for the evaluated variables (OR = 3.36; 95 % CI: 1.49–7.58). For more details, we refer to Leroy *et al*. (2008b).

Acknowledgements

The *Signal Tandmobiel*® study was the result of a collaborative effort of oral health researchers (paediatric and preventive dentistry), youth health care specialists and biostatisticians and was organized in a partnership with the Flemish dental association (Working Group Oral Health Promotion and Prevention, Flemish Dental Association) and the Flemish Association for Youth Health Care. Contributors to the project were: Dominique Declerck (Scientific co-ordinator, Catholic University Leuven), Jacques Vanobbergen (Project co-ordinator, Ghent University), Luc Martens (Ghent University), Peter Bottenberg (University Brussels), Emmanuel Lesaffre (Catholic University Leuven) and Karel Hoppenbrouwers (Catholic University Leuven). LeverElida (Unilever) financially supported data collection.

Extra support was obtained from Research Grants OT/00/35 and OT/05/60 from the Catholic University of Leuven.

The following partners collaborated in the *Smile for Life* project: Dominique Declerck (Project Coordinator) and Roos Leroy (both from the Department of Dentistry, Catholic University Leuven); Karel Hoppenbrouwers (Youth Health care at the Catholic University of Leuven and the Flemish Society for Youth Health Care); Emmanuel Lesaffre (Centre for Biostatistics, Catholic University Leuven); Stephan Vanden Broucke (Research Group for Stress, Health and Well-being at the Catholic University Leuven); Luc Martens (Dental School, Ghent University); and Erwin Van Kerschaver and Martine Debyser (Child and Family). The study was supported financially by GABA Benelux and GABA International.

References

Alaluusua, S. & Malmivirta, R. (1994) Early plaque accumulation – a sign for caries risk in young children. *Community Dent Oral Epidemiol* **22**: 273–6.

Bogaerts, K., Leroy, R., Lesaffre, E., & Declerck, D. (2002) Modeling tooth emergence data based on multivariate interval-censored data. *Statist Med* **21**: 3775–87.

Carvalho, J.C., Ekstrand, K.R., & Thylstrup, A. (1989) Dental plaque and caries on occlusal surfaces of first permanent molars in relation to stage of eruption. *J Dent Res* **68**(5): 773–9.

Carvalho, J.C., Declerck, D., & Vinckier, F. (1998) Oral health status in Belgian 3- to 5-year old children. *Clin Oral Invest* **2**: 26–30.

Cecere, S., Jara, A., & Lesaffre, E. (2006) Analyzing the emergence times of permanent teeth: an example of modeling the covariance matrix with interval-censored data. *Statistical Modelling* **6**: 337–51.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ Psycholog Meas* **XX**(1): 37–46.

Declerck, D., Leroy, R., Martens, L., *et al*. (2008) Factors associated with prevalence and severity of caries experience in preschool children. *Community Dent Oral Epidemiol* **36**:(2), 168–172.

Defranc, A., Vanden Broucke, S., Leroy, R., *et al*. (2008) Measuring oral health behaviour in Flemish health care workers: an application of the theory of planned behaviour. *Community Dental Health* **25**: 107–14.

Hausen, H. (1997) Caries prediction: state of the art. *Community Dent Oral Epidemiol* **25**: 87–96.

Klein, H. & Palmer, C.E. (1938) Studies on dental caries: I. Dental status and dental needs of elementary school children. *Public Health Reports* **53**: 751–65.

Leroy, R., Bogaerts, K., Lesaffre, E., & Declerck, D. (2003a) The emergence of permanent teeth in Flemish children. *Community Dent Oral Epidemiol* **31**: 30–9.

Leroy, R., Bogaerts, K., Lesaffre, E., & Declerck, D. (2003b) Impact of caries experience in the deciduous molars on the emergence of the successors. *EurJ Oral Sci* **111**: 106–10.

Leroy R, Bogaerts K, Lesaffre E, Declerck D (2005) Effect of caries experience in primary molars on cavity formation in the adjacent permanent first molar. *Caries Res*; **39**: 342–349.

Leroy R, Cecere S, Lesaffre E, Declerck D. (2008a) Variability in permanent tooth emergence sequences in Flemish children. *Eur J Oral Sci* **116**: 11–17.

Leroy, R., Hoppenbrouwers, K., Jara, A., & Declerck, D. (2008) Parental smoking behaviour and caries experience in preschool children. *Community Dent Oral Epidemiol* **36**(3), 249–257.

Lesaffre, E., Mwalili, S.M., & Declerck, D. (2004) Analysis of caries experience taking inter-observer bias and variability into account. *J Dent Res* **83**: 951–5.

Magder, L.S. & Hughes, J.P. (1997) Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol* **146**: 195–203.

Mühleman, R. & Son, S. (1971) Gingival sulcus bleeding – a leading symptom in initial gingivitis. *Helvetia Odontologica Acta* **15**: 105–13.

Persson, L.A. & Carlgren, G. (1984) Measuring children's diets: evaluation of dietary assessment techniques in infancy and childhood. *Int J Epidemiol* **13**(4): 506–17.

Pitts, N.B., Evans, D.J., & Pine, C.M. (1997) British Association for the Study of Community Dentistry (BASCD) diagnostic criteria for caries prevalence surveys – 1996/1997. *Community Dent Health* **14** (Suppl 1): 6–9.

Silness, J. & Loë, H. (1964) Periodontal disease in pregnancy II. Correlation between oral hygiene and periodontal condition. *Acta Odontologica Scandinavica* **22**: 121–35.

Slade, G.D. & Caplan, D.J. (2000) Impact of analytic conventions on outcome measures in two longitudinal studies of dental caries. *Community Dent Oral Epidemiol* **28**: 202–10.

Vanobbergen, J., Martens, L., Lesaffre, E., & Declerck, D. (2000) The Signal Tandmobiel® project, a longitudinal intervention health promotion study in Flanders (Begium): baseline and first year results. *Eur J Paediatr Dent* **2**: 87–96.

Vanobbergen, J, Martens, L., & Declerck, D. (2001a) Caries prevalence in Belgian children: a review. *Int J Paediatr Dent* **11**: 164–70.

Vanobbergen, J, Martens, L., Lesaffre, E., Bogaerts, K., & Declerck, D. (2001b) Assessing risk indicators for dental caries in the primary dentition. *Community Dent Oral Epidemiol* **29**: 424–34.

Vanobbergen, J., Martens, L., Lesaffre, E., Bogaerts, K., & Declerck, D. (2001c) The value of a baseline caries risk assessment model in the primary dentition for the prediction of caries incidence in the permanent dentition. *Caries Res* **35**: 442–50.

20

Subantimicrobial-dose doxycycline effects on alveolar bone loss in postmenopausal women: example of a randomized controlled clinical trial

Julie A. Stoner and Jeffrey B. Payne

20.1 Introduction

Chapter 19 summarized the design, conduct and analysis of two oral health epidemiologic studies. In contrast to those observational studies, this chapter summarizes a completed periodontitis clinical trial investigating the efficacy and safety of a specific intervention through an experimental study. The design, conduct and analysis of longitudinal periodontitis clinical trials are often complex due in part to the tooth-site-specific measures of disease that are clustered within a patient over time; differing disease progression rates across locations within a mouth and very low progression rates in periodontal maintenance patients; and multiple periodontal endpoints including clinical periodontal and oral radiographic measures, as surrogate endpoints for tooth loss, and biochemical measures. This chapter utilizes data

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5

from a clinical trial investigating the efficacy and safety of subantimicrobial dose doxycycline (SDD) in slowing the progression of oral bone loss and periodontitis in postmenopausal, osteopenic, estrogen-deficient women with moderate to severe chronic periodontitis to illustrate approaches for addressing these design and analysis complexities. Details of the study design, conduct, analysis, and results can be found in the primary study publications describing the radiographic endpoint results (Payne *et al*., 2007), the clinical endpoint results (Reinhardt *et al*., 2007), the microbiologic results (Walker *et al*., 2007) and the gingival crevicular fluid (GCF) results (Golub *et al*., 2008).

20.2 Background

Estrogen deficiency in postmenopausal women has been shown to be associated with systemic bone mineral density loss and osteoporosis (Cranney *et al*., 2002), an accelerated bone resorption rate greater than the rate of bone formation (Riggs & Melton, 1986), and an increased risk of tooth loss and oral bone loss (Tezal *et al*., 2005; Payne *et al*., 1999). Tetracyclines and their chemically modified non-antibacterial analogs have been shown to inhibit the activity of collagenase, a host-derived tissue-destructive matrix metalloproteinase, and enhance osteoblast activity, collagen production, and bone formation (Sasaki *et al*., 1992; Bain *et al*., 1997; Craig *et al*., 1998; Golub *et al*., 1983, 1999), These agents also have been shown to improve clinical attachment levels and probing depths and lower inflammation in general populations with periodontitis when used as an adjunct to scaling/root planing (Caton *et al*., 2000; Lee *et al*., 2004; Preshaw *et al*., 2004; Gurkan *et al*., 2005). Based on this background, a randomized, controlled clinical trial was designed to test the hypothesis that subantimicrobial dose doxycycline is effective and safe in slowing the progression of oral bone loss and the progression of periodontitis in postmenopausal, osteopenic, estrogen-deficient women on periodontal maintenance therapy for moderate to advanced chronic periodontitis. This project was supported by Grant Number R01DE012872 from the National Institute of Dental & Craniofacial Research (Dr. Jeffrey B. Payne, PI).The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Dental & Craniofacial Research or the National Institutes of Health.

20.3 Specific aims

The primary aim of the clinical trial was to determine whether a two-year continuous regimen of SDD (20 mg doxycycline hyclate twice daily) can reduce radiological evidence of significant alveolar bone density loss and alveolar bone height loss in postmenopausal, osteopenic, estrogen-deficient women on periodontal maintenance therapy for moderate to severe chronic periodontitis. Secondary aims were: (1) to compare the effect of placebo and SDD on clinical measures including relative clinical attachment level (RCAL), manual probing depth, bleeding on

probing, and plaque; (2) to compare the effect of placebo and SDD on collagenase activity and bone resorption markers in GCF and bone formation and bone resorption markers in serum; (3) to compare the effect of placebo and SDD on systemic bone mineral density changes (lumbar spine and femoral neck); and (4) to compare the adverse event profiles, including microbiologic antibiotic resistance, between SDD and placebo. The oral radiographic and clinical endpoints considered in this study are surrogate endpoints of tooth loss. The estimated treatment effect using surrogate endpoints is useful to the extent that the effect of the intervention on the surrogate endpoint predicts the effect on the clinical outcome, tooth loss (Fleming & DeMets, 1996). However, a study of the effect of SDD on tooth loss would require thousands of subjects followed for a long period of time, and development of strict criteria for tooth extraction. There is no agreement as to the optimal surrogate endpoint for tooth loss. Therefore, inclusion of multiple endpoints is attractive. A subset of these aims, and corresponding endpoints, will be discussed in this chapter as illustrations of some of the design and analysis complexities of periodontitis clinical trials.

20.4 Endpoints

20.4.1 Primary radiographic endpoint

The primary endpoint used to justify the size of the trial was alveolar bone density measured by the computer-assisted densitometric image analysis (CADIA) method (Payne *et al*., 1999). CADIA measures were recorded at 12- and 24-months relative to baseline. Measurements were made at the crestal and subcrestal locations of interproximal sites for 4 posterior teeth (distal site on each first premolar, distal and mesial sites on the second premolar and the first molar and mesial site on each second molar) in each quadrant (see Figure 20.1), resulting in 48 tooth/site/location measurements for a subject with all posterior teeth present (excluding third molars) at each time point of measurement.

Progression of periodontitis is tooth-site specific with some sites within a mouth demonstrating progression and some demonstrating improvement, while others remain essentially unchanged. In settings with fairly stable disease (e.g. such as subjects in this trial undergoing periodontal maintenance, which is the standard of care for patients with periodontitis and must be provided to subjects enrolled in longitudinal periodontitis clinical trials), most periodontal sites show very little change over time. Estimated treatment effects using mean change values may be very small due to the influence of the majority of sites that are stable and may be difficult to interpret (Hujoel *et al*., 1993). To address this limitation, a tooth-site level ordinal endpoint was defined with three categories reflecting gain (improvement) in alveolar bone density, stability, or decrease in alveolar bone density over time beyond a specific threshold and treatment effects were summarized using proportions, such as the proportion of sites demonstrating alveolar bone density loss. To avoid biasing the treatment group comparison, the thresholds used to define categories of progression and improvement in disease should be pre-specified, and ideally based

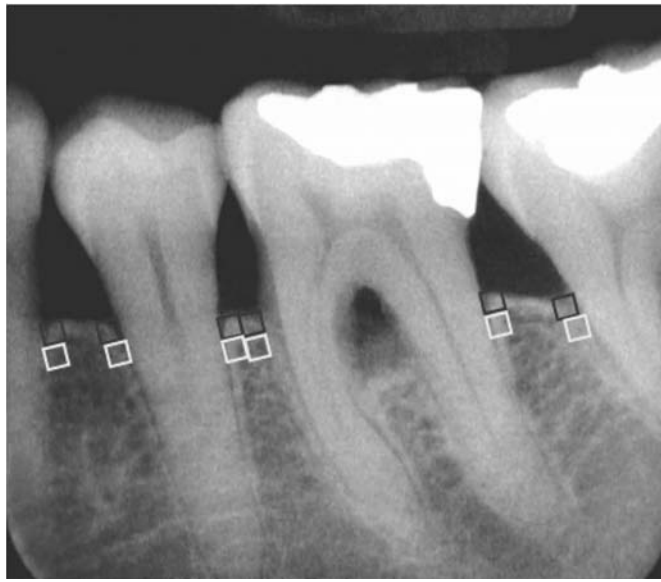


Figure 20.1 Image of crestal (1 mm × 1 mm upper black box) and subcrestal (1 mm × 1 mm lower white box) locations for radiographic alveolar bone density measurements.

on clinically significant changes (i.e. changes that are important clinically that may alter how a patient is treated or managed) (Greenstein, 2003). However, thresholds of clinically significant changes are not always available for specific populations or measurements of interest.

In this clinical trial, the cut-points for defining the categorical measures of change were based on the degree of repeatability of the radiographic measures, since no clinically significant thresholds had been established at the time of the study analysis. Specifically, the defined thresholds were based on 2 times the standard deviation of replicate measures, rounded to the nearest whole number. Repeated CADIA measurements were made on a randomly chosen sample of roughly 10 % of the study subjects to assess measurement reliability and the standard deviation of replicate measures was calculated (Osborn *et al*., 1992). The standard deviation of replicate measurements for the CADIA measures was 8.6 for the crestal location and 7.0 for the subcrestal location.The thresholds defining changes at the site level for the CADIA measure were values $\geq$ 17 at a crestal site or $\geq$ 14 at a subcrestal site defining improvement (i.e. gain in alveolar bone density), values between -17 and 17 at a crestal site and between -14 and 14 at a subcrestal site defining no change, and values $\leq$ -17 at a crestal site or $\leq$ -14 at a subcrestal site defining progression (i.e. loss of alveolar bone density). For example, if the CADIA measure, relative to baseline, was -18 at the crestal site and -10 at the subcrestal site for the mesial location of tooth number 3, this location would be categorized as demonstrating alveolar bone density loss. A location with a crestal CADIA measure of 5 and

a subcrestal CADIA measure of 7 would be categorized as stable. Tooth sites that demonstrated bone gain at one location, for example at the crestal location, and bone loss at another location, for example at the subcrestal location, were categorized as sites with bone gain. Such a discrepancy occurred in only 0.05 % of sites, however.

20.4.2 Secondary clinical endpoint

One of the clinical outcome measures was the RCAL measurement, which was made with the Florida Disk Probe using 20 g of force (Florida Probe Corp., Gainesville, FL, USA). RCAL represents a measurement from the occlusal surface to the base of the periodontal pocket and was recorded to the nearest 0.2 mm. RCAL measures were made at 4 locations (mesiobuccal, distobuccal, mesiolingual, distolingual) on each first and second premolar and each first molar and 2 sites (mesiobuccal and mesiolingual) on each second molar (see Figure 20.2). This resulted in a total of 56 measurements for a subject with all posterior teeth present at a measurement time point. Measurements were made at baseline and every 6 months during the 24-month follow-up period.

Similar to the radiographic measures, cut-points for defining the categorical measures of RCAL change, for example, over the 24-month treatment period relative to baseline, were based on the degree of repeatability. The standard deviation of replicate measurements for the RCAL measures was 0.69 mm. The changes in RCAL from baseline were each divided into 3 categories (improvement, no change, and disease progression) using thresholds defined by two standard deviations of replicate measures, rounded up to the nearest 0.5 mm. The thresholds defining significant changes at the site level for the Florida Probe RCAL change were $\leq$ -1.5 mm defining improvement (i.e., gain in RCAL), $\geq$ 1.5 mm defining progression, and values in between thresholds defining no change.

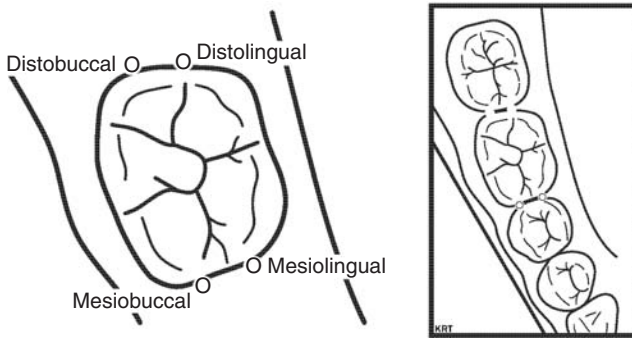


Figure 20.2 The sites for relative clinical attachment level (RCAL) measurements are diagrammatically shown on the right mandibular first molar. RCAL measurements were made so that the probe tip was placed as close to the interproximal area as possible.

20.4.3 Safety endpoint

Safety endpoints included measures of microbiologic resistance to doxycycline and other antibiotics. One specific measure that will be discussed further in this chapter is the total anaerobic count obtained on TSBA-HK (trypticase-soy blood agar supplemented with hemin and menadione) without doxycycline. Microbiological samples were collected at two posterior sites with periodontal probing depth (PPD) $\geq$ 5 mm on posterior teeth at baseline and 24-months for microbiologic analyses. The samples were pooled for microbiologic analyses and run in duplicate, resulting in two baseline and two, 24-month measures for each subject (the same sites were sampled at baseline and 24 months). The sites selected for microbiology sampling were separate from those used for the collection of gingival crevicular fluid, so that the collection of one type of sample would not have an effect on the collection of the other type of sample. If two sites with PPD $\geq$ 5 mm, that had not been sampled for gingival crevicular fluid, were not available, sites with shallower pockets were chosen. Across the study subjects at baseline, for microbial analyses, 52 % had two sites sampled with PPD $\geq$ 5 mm, 34 % had one site with PPD $\geq$ 5 mm, and 14 % had shallower (<5 mm) pockets sampled.

20.5 Study design and methods

20.5.1 Eligibility criteria

Eligible subjects were 45–70 years of age at telephone screening, were postmenopausal for at least 6 months and not receiving hormone replacement therapy (HRT), were osteopenic at either the lumbar spine or femoral neck (bone mineral density (BMD) T-score between -1.0 and -2.5, inclusive), were undergoing periodontal maintenance therapy for generalized moderate to advanced chronic periodontitis, and provided consent. Subjects were recruited at two clinical sites, the University of Nebraska Medical Center College of Dentistry, Lincoln, NE (UNMC) and the School of Dental Medicine at Stony Brook University, Stony Brook, NY (Stony Brook). All eligible subjects had at least 9 posterior teeth and at least two sites with $\geq$ 5 mm probing depths, $\geq$ 5 mm clinical attachment loss and bleeding on probing. Subjects were excluded if they had an allergy or hypersensitivity to tetracyclines; diseases or regular drug therapy that would affect the inflammatory or immune response (e.g. chronic use of non-steroidal anti-inflammatory drugs [NSAIDs]) or bone remodeling (e.g. prescription estrogens, bisphosphonates, calcitonin, and steroids); requirement for antibiotic premedication; diabetes; active periodontal therapy within the past year; normal BMD at both the lumbar spine and femoral neck (T-score above -1.0) or osteoporosis of the lumbar spine or femoral neck (T-score less than -2.5). The study protocol was approved by the University of Nebraska Medical Center Institutional Review Board and the Stony Brook Institutional Review Board (IRB).

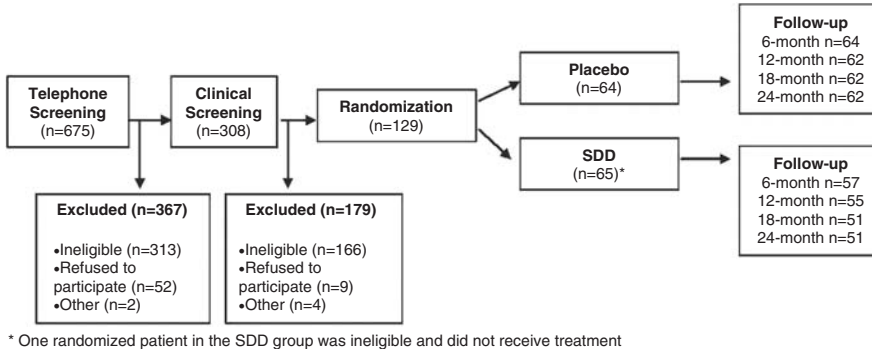


Figure 20.3 Summary of subject recruitment and follow-up.

20.5.2 Subject recruitment and follow-up

Subjects were recruited from Nebraska and Long Island, New York private periodontal and general dental practices, dental school patient pools, and advertisements. Potential subjects were sent letters describing the study and were asked to call the participating study centers. Other subjects responded to advertisements or were directly referred to clinical centers for study screening. Eligibility criteria were initially assessed through a telephone screen, occurring between June 2002 and October 2003 (see Figure 20.3). Telephone screen-eligible subjects were invited for a clinical screening visit. All subjects signed the IRB-approved consent forms at this clinical screening visit. Eligible subjects who provided consent then were randomized between June 2002 and October 2003, and subject follow-up was completed in October 2005. A summary of the timing of data measurement over the follow-up period is presented in Table 20.1.

20.5.3 Treatment assignment

Subjects were randomized in a 1:1 ratio between two treatment arms: 20 mg doxycycline twice daily (low dose or subantimicrobial dose-doxycycline; SDD) or a placebo look-alike twice daily. As the standard of care for postmenopausal women, all subjects received calcium and vitamin D supplements (1200 mg of calcium and 400 I.U. of vitamin D daily) and all subjects received periodontal maintenance every 3-4 months throughout the clinical trial, delivered by the subjects' own dental care providers and not by the study clinicians, and provided at no cost to the subjects. A computer-generated randomization list was prepared using a randomly-permuted block schema, with block size varying randomly among 4, 6, and 8. The randomization was stratified based on study center (UNMC or Stony Brook) and current smoking status (current smoker or not current smoker). To avoid potential bias, treatment assignment was blinded from the subject and all study investigators,

Table 20.1 Summary of timing of data measurement.

Assessment	Clinical screening (Day -30 to Day 0)	Baseline (Day 0)	6-month (± 1 week)	12-month (± 2 weeks)	18-month (± 3 weeks)	24-month (± 4 weeks)
Inclusion/exclusion	X					
Randomization		X				
Dispensed medications		X	X	X	X	
Medical history	X	X	X	X	X	X
Concomitant medications	X	X	X	X	X	X
Adverse events		X	X	X	X	X
Bite-wings and DEXA scans		X		X		X
Serum sample		X		X		X
Oral soft tissue exam	X	X	X	X	X	X
Tooth count and dental history		X	X	X	X	X
Gingival crevicular fluid samples		X		X		X
Oral microbiological (plaque) samples		X				X
Relative clinical attachment level and plaque		X	X	X	X	X
Bleeding on probing and probing depths	X	X	X	X	X	X
Study medication adherence assessment			X	X	X	X

except for the statistician who generated the randomization sequence and was not involved in the data analysis, and a UNMC research pharmacist who labeled the drug bottles.

20.5.4 Sample size justification

The primary endpoint used to justify the size of the trial was the CADIA measure of alveolar bone density. While the analysis focussed on an ordinal endpoint with three levels of disease progression (improved density, stable density, and loss of density), the sample size justification was simplified to only address evidence of alveolar bone density loss defined as a decrease of at least two times the standard deviation of replicate measures in alveolar bone density at the site-level from baseline.

The null hypothesis was that the proportion of sites demonstrating loss of alveolar bone density did not differ between placebo and SDD subjects and the alternative hypothesis was that the proportion of sites demonstrating loss of alveolar bone density did differ between placebo and SDD subjects. Based on our unpublished pilot study of six patients using CADIA, the expected proportion of sites with significant alveolar bone density loss at either the crestal or subcrestal region was 0.07 for the SDD group and 0.14 for the placebo group. The average number of observations to be collected at each time-point was estimated to be 18, corresponding to the measured sites of the posterior teeth assuming some tooth loss in this population. Post-baseline follow-up information on the primary endpoint was collected at 12-months and at 24-months. The sample size calculation was adjusted for an average correlation among observations sampled within a given subject's mouth and across longitudinal observations of 0.14. The sample size justification and software developed by Rochon, based on Generalized Estimating Equations (GEE) analysis, were used (1998).

The target randomization was 51 subjects per treatment group, which resulted in approximately 80 % power to detect a true difference between a probability of alveolar bone density loss of 14 % under placebo and 7 % under SDD. A drop-out rate of 20 % by the 24-month visit was accounted for by increasing the total number of randomized subjects to 128, or 64 per treatment group, resulting in 2304 site-level measurements (an average of 18 measures for each of 128 subjects) at each time point. A drop-out rate of up to 20 % was assumed in the planning stage of the trial based on the investigators' experience with the targeted population of postmenopausal women (Reinhardt *et al*., 1999). Dropouts from clinical trials decrease estimating efficiency, but may also lead to bias if the characteristics and outcomes of subjects who drop out from the study differ from those who remain, and therefore, the dropout rate should be minimized.

20.5.5 Statistical analysis methods

20.5.5.1 Primary and secondary endpoint analyses

Descriptive statistics, including percentages for categorical measures and means and standard deviation for continuous measures, were calculated (see Chapter 10). For continuous tooth-site level measures, the standard deviation values were estimated using a linear mixed model with random subject and tooth effects to account for the correlation among observations on the same subject (see Chapters 11 and 13). The CADIA values and the change in RCAL from baseline were each coded into three categories of change (improvement, no change, disease progression) using thresholds of two times the standard deviation of replicate measures, as described in the previous sections. An ordinal endpoint, with 3 categories, was chosen over a binary endpoint to capture information about disease improvement and disease progression. To account for the correlation among measures within a mouth over time, generalized estimating equations (GEE) methodology was used to fit cumulative logistic regression models for the categorical responses to compare the odds of more progressive disease (among the ordered categories of improvement, no change, and progression) over the treatment period between the SDD and placebo groups (Liang & Zeger, 1986; McCullagh & Nelder, 1989). A cumulative logistic model was chosen since an ordinal outcome variable was specified. Covariates in the regression models included time, treatment, and their interaction, where nonsignificant interaction terms were dropped from subsequent models and average treatment effects across both follow-up time points were reported unless otherwise noted. Randomization stratification factors (baseline smoking status and study center) also were included in the regression models as was the baseline outcome measurement for the RCAL measures (CADIA represents a relative change from baseline; therefore, there is no baseline measure). If only two response categories were specified, a logistic regression model was fit and if continuous measures of change were analyzed, a linear regression model was fit using a similar approach as for the cumulative logistic models.

The cumulative logistic model involves a series of modeling equations. The cumulative logistic models that were fit, without the time by treatment interaction terms, can be written as follows:

$$log\left(\frac{\pi_3}{\pi_2+\pi_1}\right) = \beta_{01} + \beta_1 SDD + \beta_2 Center + \beta_3 Time + \beta_4 Smoking,$$

$$log\left(\frac{\pi_2+\pi_3}{\pi_1}\right) = \beta_{02} + \beta_1 SDD + \beta_2 Center + \beta_3 Time + \beta_4 Smoking,$$

where β_{01} and β_{02} are intercept terms, β_i for $i = 1, \ldots, 4$ are the regression coefficients for covariates indexed by i, π_1 is the probability of disease improvement given the covariate values, π_2 is the probability of stable disease given the covariate values, and π_3 is the probability of disease progression given the covariate values. See Chapters 11 and 13 for a discussion of model fitting and interpretation. All statistical analyses were performed using SAS (version 9.1, SAS Institute Inc., Cary, NC, USA).

20.5.5.2 Safety endpoint analysis

For the microbiological safety analyses, there was interest in establishing that microbiologic resistance to multiple antibiotics, including doxycycline, for subjects treated with SDD was similar to that of placebo. This is an equivalency hypothesis. Estimation, in particular estimation of confidence intervals, is of greater use in equivalency settings than hypothesis testing (Blackwelder, 1982). The objective is to estimate a range of reasonable estimates for the parameter of interest (e.g. the difference in the mean total anaerobic count between SDD and placebo), and establish that this interval does not include clinically-significant differences. The GEE method was used to fit linear regression models for the log_{10} transformed anaerobic count. Covariates included in the model were similar to those described for the CADIA and RCAL endpoints. A 95 % confidence interval was estimated for the regression coefficient for the interaction between SDD and time, corresponding to the estimated difference in mean change in log_{10} anaerobic counts between SDD and placebo over time.

20.6 Study implementation and conduct

20.6.1 Planning and training

A detailed manual of operations was created before the study began enrolling subjects to ensure that the study procedures were implemented consistently between the study centers and across subjects over time. The manual also helped to ensure that high quality data were collected and computerized consistently. Investigator training and calibration exercises were performed at UNMC prior to enrolling any study subjects. The training focused on procedures to ensure reliable clinical and radiographic measurements (throughout the clinical trial, one examiner at each clinical center made all periodontal clinical measurements and one examiner at each clinical center took the radiographs, to eliminate inter-examiner error at each clinical center; also, one examiner, Dr. Pirkka Nummikoski, made all oral radiographic alveolar bone density and alveolar bone height measurements, in a blinded fashion, at the Longitudinal Radiographic Assessment Facility in the University of Texas Health Science Center at San Antonio), and also included a general discussion and overview of the study treatment protocol and procedures, including data collection and case report forms, with the clinical research team from both study centers (e.g., clinical examiners, clinical research assistants, and research technologists).

20.6.2 Data collection and management

Study-specific case report forms were created. The response items were coded, where appropriate, to ensure consistent reporting and to facilitate statistical analysis. Daily diaries were used by subjects to record information about adverse events and concomitant (i.e. non-study) medications and this information was transferred to

case report forms during each 6-month study visit. Subjects were asked to record all adverse events, regardless of attribution.

The case report forms were created as scannable forms, using TeleForm software by Cardiff (Vista, CA, USA), to avoid errors in hand data entry. All scanned data items were reviewed to ensure that hand-written numbers or letters were not misinterpreted by the software and data were exported to a Microsoft Access database. Data validity checks were coded based on reasonable ranges of values and data cross-checks between and within forms over time within a subject. Examples include ensuring that total tooth count did not increase over time, ensuring that completed study visits were properly ordered by calendar time (for example, checking that a reported six-month visit occurred after the baseline visit), and checking that probing depth data were not coded for a tooth after the tooth was extracted. Query reports were generated based on these validity rules and sent to the institutions for clarification of the questionable data values.

20.7 Results

A subset of the results from the clinical trial is presented in this section to illustrate some of the complexities of the data analysis for this clinical trial. The complexities highlighted in this section include the choice of summary statistics and parameter values in settings with stable disease, differential disease progression and improvement rates within a mouth, confounding baseline disease status in longitudinal assessments, and equivalency testing.

20.7.1 Primary endpoint

CADIA values at 12-months and 24-months are summarized for each treatment group in Table 20.2. Most sites demonstrated stable disease over time. For example, among the placebo subjects, 87.5 % and 81.0 % of sites at 12- and 24-months demonstrated no change beyond the specified thresholds. Descriptively, it appears that the probability of disease progression (decreases in density) is increased over time, with slightly higher rates among the placebo subjects, and that the probability of an improvement (an increase in density) is more similar between the treatment groups.

Categorized changes in CADIA measures over time are more easily interpretable in this population with stable disease than are mean change values. For example, mean CADIA values at 12-months were -0.77 (standard deviation 8.73) for the placebo arm and -0.045 (8.81) for the SDD arm and were -2.08 (10.57) for the placebo arm and -1.04 (9.82) for the SDD arm at 24-months. See Chapter 10 for a summary of descriptive statistical methods. Based on a linear regression model after adjustment for the effect of time, smoking status, and study center, we estimate that the mean CADIA value is 0.99 units greater for subjects receiving SDD relative to placebo (95 % confidence interval: -0.045 to 2.02, P-value = 0.06). These very small mean values are influenced by the majority of CADIA (change) values that were essentially 0 and are difficult to interpret clinically.

Table 20.2 Summary of CADIA measures by treatment group and study time point. Data are presented as the number and percentage of tooth sites across subjects in each treatment group demonstrating changes.

Time point of change	Change in CADIA at the tooth-site level					
	Placebo			SDD		
	Decrease in density	No change	Increase in density	Decrease in density	No change	Increase in density
12-month	91 (7.8%)	1025 (87.5%)	56 (4.8%)	72 (6.7%)	960 (88.3%)	55 (5.1%)
24-month	151 (13.0%)	939 (81.0%)	69 (6.0%)	100 (10.1%)	837 (84.2%)	57 (5.7%)

A cumulative logistic regression model was used to model the odds of more progressive disease (along ordered categories of improvement in density, no change, and decrease in density) associated with SDD and placebo treatment. Based on the models, there was no significant evidence that the effect of treatment differs over time (P-value = 0.5 time by treatment interaction). After adjustment for the effect of time, smoking status, and study center, we estimate that the odds of more progressive disease based on CADIA measures are 16 % lower for subjects receiving SDD relative to placebo, which is not statistically significant (OR = 0.84 (SDD relative to placebo), 95 % confidence interval: 0.65 to 1.08, P-value = 0.2).

20.7.2 Secondary endpoint

Baseline RCAL values for subjects are summarized for each treatment group in Table 20.3. The standard deviation values were estimated using a linear mixed model with random subject and tooth effects to account for the correlation among observations on the same subject (see Chapters 11 and 13 for additional information). The mean RCAL measures are slightly higher for the SDD group. Imbalances between treatment groups in baseline outcome measures are often important to adjust for in regression modeling. In this case, the SDD group has slightly higher

Table 20.3 Summary of baseline RCAL measures by treatment group. Descriptive statistics were calculated for the individual tooth-site level data. A linear mixed model was used to calculate the standard deviation (Std. Dev.) of measures while accounting for the nested data structure.

Study drug	Baseline RCAL measurement (mm)							
	n*	Min	25^{th} percentile	Median	75^{th} percentile	Max	Mean	Std. dev.
Placebo	3047	3.60	7.80	8.60	9.80	20.00	8.86	1.68
SDD	3156	5.40	8.00	9.00	10.00	19.80	9.11	1.69

*n represents the number of tooth sites

measures of disease on average and, therefore, may have a greater potential for larger improvements (or reductions) in RCAL over time.

Categorical changes in RCAL are summarized for each treatment group by each time point in Table 20.4. A high percentage of sites (greater than 91 % at each time point for each treatment group) in this maintenance population showed lack of disease progression or improvement based on two times the standard deviation of replicate measures. Descriptively, it appears that the risk of disease progression based on the RCAL measure is slightly lower for SDD than for placebo for visits 12-months and 18-months. The probability of improvement in RCAL is higher for SDD than for placebo for all study visits.

Site specificity of periodontitis also was observed; all sites were not equally susceptible to periodontitis progression. The sites between the first and second molars were more likely to demonstrate disease progression, based on descriptive statistics. For example, at 24 months among the placebo subjects, 3.0 % of first molar-second molar interproximal sites demonstrated progression compared to 1.8 % of sites between second pre-molars and first molars and 2.0 % of sites between first and second pre-molars.

Based on regression modeling, there is no significant evidence that the effect of study treatment differs over time (P-value = 0.2, time by treatment interaction). Without adjustment for confounding or design factors, including baseline smoking status, study center, and baseline RCAL measures, the odds of disease progression based on RCAL measures are 28 % lower for subjects receiving SDD relative to placebo, which is statistically significant (OR = 0.72 (SDD relative to placebo), 95 % confidence interval: 0.59 to 0.87, P-value = 0.0009). After adjustment for the effect of time, smoking status, study center, and baseline RCAL measurements, we estimate that the odds of disease progression based on RCAL measures are 19 % lower for subjects receiving SDD relative to placebo, which is statistically significant (OR = 0.81 (SDD relative to placebo), 95 % confidence interval: 0.67 to 0.97, P-value = 0.03). Similar estimates are found when the model includes only a treatment term and a baseline RCAL term, suggesting that baseline RCAL is an important confounding factor. Based on odds ratio estimates, the treatment effect size is modest.

20.7.3 Safety endpoint

Microbiologic samples were analyzed in duplicate at baseline and at the 24-month visit. Table 20.5 summarizes the total anaerobic counts at the baseline and 24-month visits for each study drug group. The baseline and 24-month values summarize all duplicate measurements made on each subject at each time point. The counts have been transformed by the log_{10} function.

Based on the descriptive summary, it appears that the baseline and 24-month distributions were similar between the study drug groups. For both drug groups, the log_{10} anaerobic count values do not appear to change very much over the follow-up period.

Table 20.4 Summary of changes in RCAL measures by treatment group and study time point. Data are presented as the number and percentage of tooth sites across subjects in each treatment group demonstrating changes.

Time point of change	Change in RCAL at the tooth-site level					
	Placebo			SDD		
	Improvement	No change	Disease progression	Improvement	No change	Disease progression
6-month	77 (2.6%)	2788 (95.0%)	69 (2.4%)	123 (4.5%)	2583 (93.6%)	54 (2.0%)
12-month	107 (3.8%)	2650 (93.3%)	83 (2.9%)	126 (4.7%)	2496 (93.5%)	48 (1.8%)
18-month	95 (3.4%)	2655 (93.6%)	86 (3.0%)	139 (5.6%)	2294 (92.3%)	52 (2.1%)
24-month	108 (3.9%)	2587 (93.4%)	74 (2.7%)	129 (5.2%)	2268 (91.8%)	75 (3.0%)

Table 20.5 Summary of Log_{10} anaerobic count measures by treatment group and study time point. Samples were processed in duplicate and data are summarized at the duplicate microbiologic sample level.

Study drug and time point		n*	Log_{10} anaerobic count				
			Min	25^{th} percentile	Median	75^{th} percentile	Max
Placebo							
	Baseline	126	5.86	6.79	6.94	7.08	7.68
	24-month	116	4.43	6.74	6.91	7.15	7.73
SDD							
	Baseline	128	6.34	6.82	7.00	7.15	7.48
	24-month	102	6.34	6.78	6.99	7.15	7.45

*n represents the number of measurements across samples processed in duplicate.

Based on a GEE regression analysis of the log_{10} anaerobic count values, the mean change in log_{10} anaerobic counts is estimated to be a 0.04 unit decrease over the two-year treatment period for both placebo and SDD subjects. There is no significant evidence that the change over time in the mean log_{10} anaerobic counts differs between subjects assigned to SDD relative to subjects assigned to placebo (P-value = 0.97). Assuming a two-sided alpha level of 0.05, it is reasonable to expect the difference in the change in the mean log_{10} anaerobic counts over time to be from 0.16 units less to 0.16 units greater for subjects assigned to SDD relative to subjects assigned to placebo. While the point estimate of the difference in mean change values between the treatment groups is 0, the confidence interval for the difference in means is rather wide, where the value of 0.16 is 4 times the average change in the placebo group (0.04). The interval width is driven in part by the variability in the estimate of the mean change and the limited sample size. The sample size was justified relative to the primary radiographic endpoint (alveolar bone density) and was not sufficient to derive tight estimation intervals for the secondary, microbiologic endpoints. So, while there is no evidence of a significant difference between the treatment group mean changes over time, the limited sample size does not allow us to establish equivalence.

20.8 Concluding remarks and future directions

The trial described in this chapter highlights several complexities that arise in the design and analysis of clinical trials in periodontal research and suggest areas of future research and methodological development. Modeling procedures need to be selected carefully to ensure easily interpretable, and clinically-meaningful, results. There are multiple disease measures of interest, including radiographic measures, such as alveolar bone density and alveolar bone height; periodontal clinical measures, such as RCAL; biochemical measures such as GCF collagenase levels;

microbiologic measures such as total anaerobic counts; and systemic measures, such as serum biomarkers of bone resorption and formation and femoral neck bone mineral density. To better understand the effect of medications on patients' oral and systemic health, it is useful to collect measures for multiple types of outcome endpoints throughout the treatment and follow-up period for comparison between the intervention and control groups. This approach does, however, complicate the interpretation of the results: an example is when the estimated treatment effect is not consistent across all endpoints. Multiple hypothesis testing also increases the type I, or false positive, error rate.

The statistical methodology utilized in the analysis of this clinical trial accounted for the correlation among measurements in the same subject over time using a standard GEE method of analysis that avoids complicated modeling of the correlation structure. Periodontal data from longitudinal follow-up studies exhibit complicated nested correlation structures due to the sampling of sites within teeth within mouths over time where the number of correlated observations per person is large. In this complicated correlated data setting, there is a need to develop algorithms for implementing more efficient estimating procedures, particularly when considering ordered, categorical outcome measures where the regression modeling approaches that are implemented in standard statistical software packages offer only limited options to account for correlated data structures.

References

Bain, S., Ramamurthy, N. S., Impeduglia, T., Scolman, S., Golub, L. M. & Rubin, C. (1997) Tetracycline prevents cancellous bone loss and maintains near-normal rates of bone formation in streptozotocin-diabetic rats. *Bone* **21**, 147–53.

Blackwelder, W. C. (1982) 'Proving the null hypothesis' in clinical trials *Controlled Clinical Trials* **3**, 345–53.

Caton, J. G., Ciancio, S. G., Blieden, T. M., *et al*. (2000) Treatment with subantimicrobial dose doxycycline improves the efficacy of scaling and root planing in patients with adult periodontitis. *Journal of Periodontology* **71**, 521–32.

Craig, R. G., Yu, Z., Xu, L., *et al*. (1998) A chemically modified tetracycline inhibits streptozotocin-induced diabetic depression of skin collagen synthesis and steady-state type I procollagen mRNA. *Biochimica et Biophysica Acta* **1402**, 250–60.

Cranney, A. Guyatt, G., Griffith, L., *et al*. (2002) Meta-analyses of therapies for post-menopausal osteoporosis. IX: Summary of meta-analyses of therapies for postmenopausal osteoporosis. *Endocrine Reviews* **23**, 570–78.

Fleming, T. R. & DeMets, D. L. (1996) Surrogate end points in clinical trials: Are we being misled? *Annals of Internal Medicine*. **125**, 605–13.

Golub, L. M., Lee, H. M., Lehrer, G., *et al*. (1983) Minocycline reduces gingival collagenolytic activity during diabetes: Preliminary observations and a proposed new mechanism of action. *Journal of Periodontal Research* **18**, 516–26.

Golub, L. M., Ramamurthy, N. S., Llavaneras, A., *et al*. (1999) A chemically-modified nonantimicrobial tetracycline (CMT-8) inhibits gingival matrix metalloproteinases, periodontal breakdown, and extra-oral bone loss in ovariectomized rats. *Annals of the New York Academy of Sciences* **878**, 290–10.

Golub, L. M., Lee, H. M., Stoner, J. A., *et al*. (2008) Subantimicrobial-dose doxycycline (SDD) modulates GCF biomarkers of periodontitis in postmenopausal osteopenic women. *Journal of Periodontology* **79**, 1409–18.

Greenstein, G. (2003) Clinical versus statistical significance as they relate to the efficacy of periodontal therapy. *Journal of the American Dental Association* **134**, 583–91.

Gurkan, A., Cinarcik, S. & Huseyinov, A. (2005) Adjunctive subantimicrobial dose doxycycline: effect on clinical parameters and gingival crevicular fluid transforming growth factor-beta levels in severe generalized chronic periodontitis. *Journal of Clinical Periodontology* **32**, 244–53.

Hujoel, P. P., Baab, D. A. & DeRouen, T. A. (1993) Measures of treatment efficacy. *Journal of Clinical Periodontology* **20**, 601–5.

Lee, J. Y., Lee, Y. M., Shin, S. Y., *et al*. (2004) Effect of subantimicrobial dose doxycycline as an effective adjunct to scaling and root planing. *Journal of Periodontology* **75**, 1500–8.

Liang, K. Y. & Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman & Hall, Chapter 5

Osborn JB, Stoltenberg JL, Huso BA, Aeppli DM, & Pihlstrom BL. (1992) Comparison of measurement variability in subjects with moderate periodontitis using a conventional and constant force periodontal probe. *J Periodontology* **63**, 283–9.

Payne, J. B., Reinhardt, R. A., Nummikoski, P. V. & Patil, K. D. (1999) Longitudinal alveolar bone loss in postmenopausal osteoporotic/osteopenic women. *Osteoporosis International* **10**, 34–40.

Payne, J. B., Stoner, J. A., Nummikoski, P. V., *et al*. (2007) Subantimicrobial dose doxycycline effects on alveolar bone loss in postmenopausal women. *Journal of Clinical Periodontology* **34**, 776–87.

Preshaw, P. M., Hefti, A. F., Novak, M. J., *et al*. (2004) Subantimicrobial dose doxycycline enhances the efficacy of scaling and root planing in chronic periodontitis: a multicenter trial. *Journal of Periodontology* **75**, 1068–76.

Reinhardt, R. A., Payne, J. B., Maze, C. A., *et al*. (1999) Influence of estrogen and osteopenia/osteoporosis on clinical periodontitis in postmenopausal women. *Journal of Periodontology* **70**, 823–8.

Reinhardt, R. A., Stoner, J. A., Golub, L. M., *et al*. (2007) Efficacy of subantimicrobial dose doxycycline in postmenopausal women: clinical outcomes. *Journal of Clinical Periodontology* **34**, 768–75.

Riggs, B. L. & Melton, L. J. (1986) Involutional osteoporosis. *New England Journal of Medicine* **314**, 1676–86.

Rochon, J. (1998) Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine* **17**, 1643–58.

Sasaki, T., Ramamurthy, N. S. & Golub, L. M. (1992) Tetracycline administration increases collagen synthesis in osteoblasts of streptozotocin-induced diabetic rats: a quantitative autoradiographic study. *Calcified Tissue International* **50**, 411–19.

Tezal, M., Wactawski-Wende, J., Grossi, S. G., Dmochowski, J. & Genco, R. J. (2005) Periodontal disease and the incidence of tooth loss in postmenopausal women. *Journal of Periodontology* **76**, 1123–8.

Walker, C., Puumala, S., Golub, L. M., *et al*. (2007) Subantimicrobial-dose doxycycline effects on osteopenic bone loss: Microbiology results. *Journal of Periodontology* **78**, 1590–1601.

Index

Statistical and Methodological Aspects of Oral Health Research Edited by E. Lesaffre, J. Feine, B. Leroux, D. Declerck
 ISBN: 978-0-470-51792-5