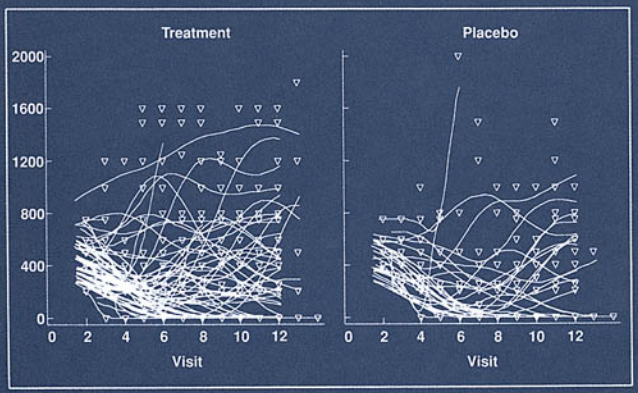
Statistical Methods for Clinical Trials



Mark X. Norleans

STATISTICAL METHODS FOR CLINICAL TRIALS

Biostatistics: A Series of References and Textbooks

Series Editor

Shein-Chung Chow

President, U.S. Operations
StatPlus, Inc.
Yardley, Pennsylvania

Adjunct Professor
Temple University
Philadelphia, Pennsylvania

1. *Design and Analysis of Animal Studies in Pharmaceutical Development*, edited by Shein-Chung Chow and Jen-pei Liu
2. *Basic Statistics and Pharmaceutical Statistical Applications*, James E. De Muth
3. *Design and Analysis of Bioavailability and Bioequivalence Studies, Second Edition, Revised and Expanded*, Shein-Chung Chow and Jen-pei Liu
4. *Meta-Analysis in Medicine and Health Policy*, edited by Dalene K. Stangl and Donald A. Berry
5. *Generalized Linear Models: A Bayesian Perspective*, edited by Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick
6. *Difference Equations with Public Health Applications*, Lemuel A. Moyé and Asha Seth Kapadia
7. *Medical Biostatistics*, Abhaya Indrayan and Sanjeev B. Sarmukaddam
8. *Statistical Methods for Clinical Trials*, Mark X. Norleans

ADDITIONAL VOLUMES IN PREPARATION

STATISTICAL METHODS FOR CLINICAL TRIALS

MARK X. NORLEANS

MARCEL DEKKER, INC. NEW YORK • BASEL

ISBN: 0-8247-0467-3

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fax: 41-61-261-8896

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

To

Naili, Jack, and Maxwell

This Page Intentionally Left Blank

Series Introduction

The primary objectives of the *Biostatistics* book series are to provide useful reference books for researchers and scientists in academia, industry, and government, and also to offer textbooks for undergraduate and/or graduate courses in the area of biostatistics. This series provides comprehensive and unified presentations of the statistical designs and analyses of important applications in biostatistics, such as those in biopharmaceuticals. A well-balanced summary will be given of current and recently developed statistical methods and interpretations for both statisticians and researchers and scientists with minimal statistical knowledge who are engaged in applied biostatistics. The series is committed to providing easy to understand, state-of-the-art reference books and textbooks. In each volume, statistical concepts and methodologies are illustrated through examples.

Clinical development in pharmaceutical research and development is a lengthy and costly process. It is necessary, however, to provide substantial evidence regarding the efficacy and safety of the pharmaceutical entity under investigation prior to regulatory approval. To ensure the success of clinical development in pharmaceutical research and development, good clinical practices (GCP) are essential. Biopharmaceutical statistics play an important role for the implementation of good clinical practices. Good statistics practices (GSP) provide a valid and fair assessment of the pharmaceutical entity under investigation with a desired accuracy and reliability.

Statistical Methods for Clinical Trials discusses important statistical concepts including statistical power, multiplicity, testing for therapeutic equivalence, and missing values, which are commonly encountered in clinical research and development. This volume also provides a unique summarization of most current statistical methodologies for analysis of longitudinal and survival data, which are commonly collected for the evaluation of the safety and efficacy of the pharmaceutical entities under investigation. In addition, this volume provides statistical interpretation of the results from various procedures of SAS, such as four types of sum of

squares in PROC GLM, repeated measures in PROC MIXED, and categorical data in PROC GENMOD.

Statistical Methods for Clinical Trials serves as a bridge among biostatisticians, clinical researchers and scientists, and regulatory agencies by providing not only a firm understanding of key statistical concepts regarding design, analysis, and interpretation of clinical studies, but also a good sense for the planning and the conduct of clinical trials in clinical research and development.

Shein-Chung Chow

Preface

Because human subjects are constantly influenced by innumerable known and unknown factors, individual patients are not directly comparable. The direct cause of this incomparability is the confounding among these factors. The greatest challenge in clinical research is the control of confounding. Unlike other statistical books, which are either filled with statistical theories that do not touch on this reality, or are too superficial in statistical technicality to be truly useful in the design and analysis of clinical studies, this book faces that challenge and presents solutions.

This book is comprehensive. It covers almost all useful statistical methods that medical researchers may come across in the design and analysis of clinical studies. This book is up-to-date and advanced. It includes frequently debated issues in the design and analysis of clinical trials, the recent development in meta-analysis and data visualization, and some of the most recent statistical methods that are frequently requested by statistical authorities in the evaluation of clinical trials. This book is innovative. It presents new ideas, new perspectives, and new techniques. This book is also controversial. It promotes graphical data analysis for data visualization as the gold standard as opposed to statistical testing; it streamlines a variety of existing statistical techniques into a single line of thinking so that medical researchers can focus on their research interest without having to be concerned with the statisticians' craft.

Like all human learning activities, clinical research requires measurement and judgment, and the two cannot be separated. The ideal candidates for the job are people who know the essential techniques for measurement and have the professional experience for judgment. Therefore, this book is primarily targeted to those who are experienced

in clinical research and desire to know more about the techniques for measurement. Research physicians, epidemiologists and medical writers will cheer this book. It will not take them long to learn all they need to know about statistics. They will be set free to do whatever analysis they desire to do without having to passively rely on what is available and deliverable from statisticians. Just as physicians order and interpret diagnostic tests without too much involvement in the technical details, the goal of this book is to empower medical researchers to design cost-efficient clinical study programs and conduct intelligent data analysis but leave most of the technical details to statisticians or personnel experienced in computation and computer programming.

But the audience is not limited to medical researchers. Data analysts, applied statisticians and statistical programmers will welcome this book. It re-exposes them to statistics from a completely different perspective. From that perspective, their thinking and analyses will come across easily to medical researchers. This book clarifies issues of multiplicity, statistical power, testing for equivalence, and missing data that may have been haunting them since their first day of statistical training. By mastering the skills in this book, they can provide valuable services to clinical research. Fellows, senior residents, senior medical students, and senior students in health sciences and applied statistics will find this book a useful source on research methodology. Students are set free from the burden of digesting the unrealistic statistical theory of Neyman and Pearson. Their focus is directed to research methodology and essential quantification techniques. By learning the skills in this book, they will be ready to participate and make immediate contributions to clinical research projects. The book provides research initiatives for students with a statistical major to study toward advanced degrees in statistics. Mathematical statisticians, who often talk about clinical studies without personal experience, may find this book challenging. This book criticizes the statistical theory of Neyman and Pearson that has been dominating current statistical educational programs, presents that theory's most adverse impact on clinical research practice, and completely removes that theory from playing a roll in clinical research. This book ignores the asymptotic theory, presents the maximum likelihood techniques in the original sense of Ronald A. Fisher, and streamlines a variety of computational techniques for the analysis of variance.

Statistics, as it is conceived today, is mostly a job for statisticians trained in academic programs, and it is, for the most part, an arcane discipline feared by most medical researchers. This book, written in plain English, will help conquer the fear. A good deal of mathematical training is not required to read and understand this book. Experience in clinical research is helpful, but not essential. No technical jargon and mathematical derivations appear in this book. Although words that have special meaning in the statistical literature may appear in this book, they are not used in that technical sense unless explicitly so defined. Critical concepts and techniques are illustrated with artificial data. I have made my best effort to keep the writing straight to the point. The pace is a bit fast, and conciseness is emphasized. References are kept to the minimum. Interested readers are referred to the books of Ronald A. Fisher published in the early decades of the twentieth century, which, in my opinion, are truly original, inspiring, and relevant to clinical research. For a list of other statistical references of uncertain relevance to clinical research, readers are referred to the dissertation of Xuemin Xie. The names of statistical authors appear only when referring to those names is conventional or necessary to avoid ambiguity.

Readers with different backgrounds may approach this book differently. Those who do not have prior exposure to statistics may read the first three chapters and become immediately productive. Chapter Eight concentrates on fundamental principles for setting up comparable groups, basic techniques for estimating sample sizes, and measures of information for comparing efficiency. That chapter is intended to help medical researchers design cost-efficient clinical study programs. Chapter Nine presents guidelines and techniques for integrated analysis of multiple related studies with focus on the quality of clinical studies and the consistency and heterogeneity of study results. Meta-analysis is discussed in the last section of Chapter Nine. Chapter Seven discusses the techniques for survival analysis. The basic measures in the first three sections are sufficient for the analysis of survival data for all practical purposes. Most of the chapters stand alone, and cross-referencing is kept to the minimum. Non-statistical readers may postpone Chapters Four, Five and Six until they really want to peek into statisticians' armamentaria. Chapter Four provides the core for understanding the analysis of variance technique. Chapter Five illustrates some of the controversies among statisticians on the use of the analysis of variance

technique. Chapter Six simply presents some technical variations of the analysis of variance, which are of little use in clinical research but tend to be overplayed by statisticians. Those who have been exposed to statistics, especially the theory of Neyman and Pearson, need to read Chapter Ten before proceeding beyond Chapter Two. Chapter Ten clearly states my views on the misuse of statistics and will help readers to understand the book better.

I hope this book lights a torch. I hope it is not just a torch for medical researchers, but a torch that lights up thousands of others carried along by researchers in other disciplines who need statistics for data presentation and unambiguous communication. What transcends all the technicalities presented in this book is the right use of human reasoning power, with which all intelligent people are equally concerned. That unleashed human reasoning power should make this book exciting to read.

Mark X. Norleans

Acknowledgment

I thank my parents Guifen Sui and Xun Xie for their faith in education. They made a great sacrifice of their personal life to support all their three children to complete the costly dental and medical education while they were struggling for food. I will always carry their hope no matter where I go. I thank Dr. Naili Duan for her love, passion and care that fill my life with water and fire. I thank the people in the state of Louisiana who generously sponsored my graduate training in the Louisiana State University Medical Center, where I started my adventure in the United States. I cherish that precious experience of both financially and intellectually free education.

Dr. William Ted Cotton, an English professor in Loyola University, New Orleans, proofread the first draft of this book and made it much more readable and presentable. Dr. Monroe Karetzky, Dr. Naili Duan, and Mr. Christopher Miller offered a peer review of the first draft. Dr. Shein-Chung Chow reviewed the first draft and suggested the addition of meta-analysis, blinding, and good clinical practice guidelines. Dr. Chow also recommended the book to Marcel Dekker, Inc. for publication. Drs. Glenn Carlson, Steven Simmonson, Kanaga Sundaram and Ms. Kimberly Williams showed great interest and encouraged the development of this book. Ms. Elaine Floyd and Dr. Alexandre Todorov helped me in contract negotiation. Belmont Research, Inc. provided the software for graphical analysis. Ms. Kathleen Baldonado assisted and coordinated the technical development of this book.

This Page Intentionally Left Blank

Contents

1
The Logical Basis of Clinical Studies and Statistical Methods

Summary

Due to the complexity of human subjects, clinical studies are subject to confounding from innumerable factors. Characterization and comparison of groups of patients, as opposed to individual patients, afford a means for the control of confounding. Therefore, groups, or comparable groups, of patients form the logical basis of clinical studies. In clinical study, factors that potentially exert effects on the responses are categorized into the controlled and uncontrolled factors. For clinical research, statistics concerns measurement for characterization of the data from groups of patients. Statistical evaluation is comparison of the effects of the controlled and uncontrolled factors. The statistical methods presented in this book include graphical data analysis, comparison of summary measures, the analysis of dispersion, and the analysis of variance.

1.1 The logical basis of clinical studies

The greatest challenge in clinical studies is the ever presence of

confounding. Confounding is logical indetermination, which happens when a claimed effect can be logically attributed to multiple possible causes. If two patients on different treatments, for instance, are also different in regard to a genetically inherited trait, as illustrated in the following table,

Trait A	Treatment A	**Patient 1**, response = 50
Trait B	Treatment B	**Patient 2**, response = 100

then the difference between the responses can not be attributed solely to the effects of treatment. This is because there is no logical reason *not* to attribute the difference to the effects of that genetic trait. When numerous causes are confounded in this manner, it is logically impossible to determine a causal relationship.

In general, clinical studies are carried out for either characterization of a disease entity or comparison of the effects of different therapeutic interventions. Due to the extreme complexity of human subjects, who are constantly influenced by innumerable known and unknown factors, clinical studies for either of the purposes cannot be based on individual patients. A disease entity characterized by observations made on an individual patient, for instance, cannot be generalized to a different patient with the same disease. This is simply because we cannot separate a disease from the patient who has the disease, and patients are all different. In other words, disease and characteristic of the patient are confounded in interpretation of clinical findings. Therefore, in clinical practice, a disease is usually documented by the common manifestations of a group of patients, with an equal appreciation of the diversity of individual patients. For the same reason, comparisons of individual patients are almost always inconclusive. For any comparison of individual patients to be conclusive, the patients must be identical in all aspects other than the few designated factors under comparison, which is clearly infeasible. What is operable and generally accepted in the current clinical research practice is comparison of groups of patients. Groups of patients can be made comparable by fairly distributing all potentially confounding factors to the groups by means of randomization, stratification, and blinding. As such, a common background is established for all the groups, and against this

background, any difference among the groups may be logically attributed to the effects of the grouping factor or factors.

Therefore, the logical basis of clinical studies is ***group of patients*** if the purpose is to characterize a disease entity or ***comparable groups of patients*** if the purpose is to compare the effects of different therapeutic interventions. Characterization or comparison of groups, as opposed to individual patients, affords a means for the control of confounding.

1.2 The principle of statistical methods

As we know, patients are constantly influenced by innumerable known and unknown factors. In clinical study, these factors are categorized into controlled factors and uncontrolled factors. The controlled factors are those whose effects are either being studied or controlled by stratification. The concept of stratification is discussed in Chapter Eight. The uncontrolled factors are all other factors that are yet to be identified or have not been controlled by stratification. For example, in a multicenter trial in which patients in each center are randomly assigned to treatment groups, treatment and center are the controlled factors, whereas other factors, including age, sex, race, baseline measures, medical history, and etc., are the uncontrolled factors. Treatment is the factor being studied, while center is the factor that stratifies the patients so that treatment effects can be compared within centers. In this sense, the effects of center on patients' responses are controlled for a more precise comparison of the treatment effects. Other factors are not controlled because the design does not provide a mechanism that precisely controls their distribution among the treatment groups. It could happen, for instance, that most of the women appear in one treatment group and most of the men in another. Then, the effects of gender would confound with the effects of treatment. Therefore, the effects of the uncontrolled factors potentially confound with the effects of the controlled factors.

In clinical research, statistics concerns measurement for characterization of data from groups of patients. Parallel to the categorization of controlled and uncontrolled factors is the categorization of statistical measures into those that quantify the effects of the controlled factors and those the overall effects of the uncontrolled

factors. Different purposes dictate the view and use of those two categories of statistical measures. Measures of the controlled factors may be viewed as the signals of interest and used to characterize a group of patients; measures of the uncontrolled factors may be viewed as the noise in the background and used to measure the degree of variation and the quality of characterization. In a study where the patients are randomly assigned to two treatment groups, for instance, the treatment is the controlled factor, and factors other than the treatment are uncontrolled. If we use the mean and standard deviation to summarize the responses of patients in each treatment group, the mean quantifies the effects of treatment, the controlled factor, and the standard deviation quantifies the effects of all other factors, the uncontrolled factors. While each mean characterizes the treatment group as a whole, its standard deviation measures the data variation from that mean and may be used to indicate the quality of that mean, as a single measure, in characterizing the responses of that group of patients.

Because the effects of the uncontrolled factors potentially confound with the effects of the controlled factors, statistical analysis of clinical data constantly involves comparison of the effects of controlled and uncontrolled factors. When music is played in a noisy stadium, it has to be loud enough to be enjoyed. Indeed, most researchers insist that the effects of the controlled factors be significant only when they are clearly distinguishable from the effects of the uncontrolled factors. While this principle is not much arguable, what is open to dispute is how these claimed effects are quantitatively measured and the extent of difference between them.

If, for instance, we use the mean to characterize treatment effect and its standard deviation to characterize the effects of the uncontrolled factors, then their ratio, mean ÷ standard deviation, of 2 implies that 50% of the treatment effects confound with the effects of the uncontrolled factors. If the ratio is 1.3, then up to 70% of the treatment effects, as measured by the mean, confound with the effects of the uncontrolled factors. However, the significance of this confounding can only be appreciated by the researchers in the context of the clinical study. If the disease entity being studies highly fluctuates, presumably caused by the uncontrolled factors, treatment effects really have to be substantial to emerge from a chaotic background by 30%. On the other

hand, if the behavior of the disease entity is fairly robust against the effects of the uncontrolled factors, the treatment effects may be trivial for practical purpose even though they are outstanding of the background by 50%.

Statistics *per se* does not make any judgment. It is false that statistics can tell us the degree of truthfulness of a conclusion with p-value; it is false that statistics can tell us how confident we can generalize the result from a single study to the general population; it is also false that statistics can give us any guideline on the adequacy of sample sizes and the 'power' of a study. Those who have pre-exposed to statistics, especially the theory of Neyman and Pearson, may read Chapter Ten for a discussion on this matter. The immediate point is that statistics concerns only measurement, and it is humans who judge and decide for their purposes.

1.3 Statistical methods for clinical studies

The following analyses of the data from a blood sugar trial present an overview of how statistical methods may be utilized for effective evaluation of clinical study and unambiguous communication. The trial was to compare the effect of drug D to that of placebo on blood sugar. Five patients were on placebo and seven on drug D, and the blood sugar values are tabulated as follows:

Treatment	Blood sugar values						
Drug D	67	123	322	232	89	109	42
Placebo	89	80	140	108	96		

The question is whether drug D and placebo have different effects on these patients' blood sugar.

1.3.1 Data visualization and graphical data analysis

The gold standard for clinical data analysis is visualization of data with graphical techniques. Graphics presents patterns without losing sight of individual data values. Patterns characterize the data and highlight the effects of the controlled factors, while individual data values fully represent the effects of the uncontrolled factors. Graphics is

highly informative and appeals to human eyes. Complex information contained in huge volume of numerical values can be condensed on a single page of graphs, and viewers can synchronize their eyes and mind to grasp the information quickly. Furthermore, multiple graphs can be cross-displayed by the factors under study to facilitate visual comparisons. Finally, graphical data presentation enhances scientific communication. The mean and standard deviation can never fully represent the rich information in clinical data, let alone p-value. Appreciation of clinical data largely depends upon the experience and purpose of the analyst. Graphics presents complete data unbiasedly and leaves judgment of the data to the viewers. Judgment is, after all, a human intellectual activity, which is so complex and, for the most part, intangible, that it is far beyond the scope of statistics.

Graphical data analysis requires fast computer and large memory. Because of the insufficient computing power and high cost of memory devices in the past, sophisticated graphical data analysis techniques were not widely available, and most researchers do not have much exposure to the concept and techniques of graphical data analysis. Chapter Two is an introduction to the garden of graphical data analysis and presents the beauty of ***CrossGraphs***®, a great graphical data analysis software package made available by Belmont Research. Nevertheless, to get the flavor now, let us look at this symmetric cumulative frequency plot of the blood sugar data:

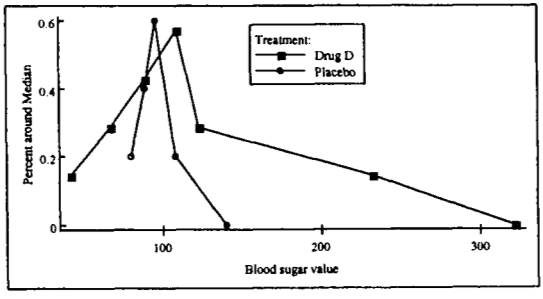


Figure 1.1 Frequency distribution of blood sugar data

® CrossGraphs is a trademark of Belmont Research, Inc.

The horizontal axis marks blood sugar values, and the vertical axis marks the percentiles for the blood sugar values less than the median and 1-percentiles for the blood sugar values greater than the median. The peak of each curve represents the data value closest to the median. Chapter Two describes the construction of symmetric cumulative frequency plot.

As we can clearly see that data distributions in the two groups overlap, and the medians are very similar. However, the data in drug D group scatter over a greater span than the data in placebo group. This difference in variation or dispersion indicates that the effect of drug D may be quite different from that of placebo. Such diverse responses to drug D suggest that some patients responded to drug D beautifully while others might not respond at all. Then, characterization of the responders and non-responders will help gain insight into the disease and pharmacodynamics of the drug. Had we compared only the medians, this important information on dispersion might have been missed.

1.3.2 Comparison of summary measures

Summary implies that few measures are being used to represent many data values. The most commonly used summary measures are the mean and median. Sometimes, few percentiles are used to characterize data distribution. Summary is attractive for its simplicity. Indeed, if the data from thousands of patients can be fully represented by the mean, then the comparison of two treatment groups is nothing more than a simple contrast between the means. The reality is, however, that patients are all different, and the uncontrolled factors will make their response to any therapeutic intervention vary greatly. Thus, the immediate question to data analysis by summary is how well the few are able to represent many?

Quality is vital for the usefulness of a summary measure. The single most important measure of quality is the number of observations. In small study with few patients, each patient makes a significant contribution to the summary measure. If an uncontrolled factor has a significant impact on a patient, that impact will largely pass through that patient and alter the magnitude of that summary measure. However, in

large study with many patients, each patient makes a relatively small contribution to the summary measure, and therefore, the effects of the uncontrolled factors on any single patient are proportionally dampened. This concept and its implications are illustrated in Chapters Three and Seven, and utilized in Chapter Eight for estimating sample size. The other two commonly used measures of quality for the mean or median are the standard deviation and average deviation. They measure the average difference between the summary measure and the actual data values and reflect the effects of the uncontrolled factors that preclude uniform responses to a therapeutic intervention from different patients. The standard deviation and average deviation are defined in Chapter Three.

Suppose the mean is chosen to summarize the data from the blood sugar trial, and the standard deviation (*std*) is chosen to measure the quality of summarization. The following table summarizes the analysis with these summary measures:

Table 1.1 A Summary of the Blood Sugar Data

Treatment	Number of Patients	Mean	Standard Deviation	Std / Mean
Drug D	7	140	93	66%
Placebo	5	102	21	21%

This study has twelve patients, with seven on drug D and five on placebo. The numbers of patients and their distribution between the groups under comparison must always be shown in the summary table. They are the most important measures for assessing the strength of evidence and determine the robustness and reliability of other summary measures. Without sufficient quantity of observations, nothing else matters. The mean blood sugar for Drug D is 140, and placebo 102. The standard deviation measures the mean distance between the mean and the actual blood sugar values. On average, the mean of drug D group deviates from the actual data by 93, and placebo by 21. However, since the magnitude of standard deviation depends upon the magnitude of the mean, a more appropriate measure of quality for the mean is the ratio of standard deviation and mean, better called percent deviation. For the blood sugar data, the mean appears to be a good summary for the

placebo group because, on average, the data deviate from the mean by only 21%. However, the mean does not seem to represent drug D group satisfactorily because, on average, the mean deviates from the data by as much as 66%. Thus, the mean of 140 conveys little information on the actual blood sugar values in drug D group, and consequently, comparison of means between the two groups has no meaning.

1.3.3 The analysis of variance

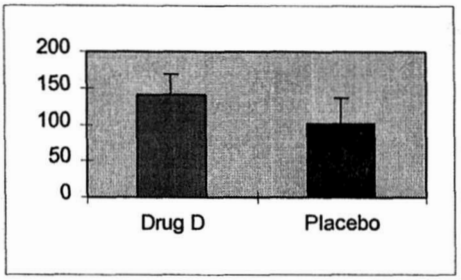
The analysis of variance technique is entirely due to Ronald A. Fisher. It is truly a useful tool for data analysis, especially for simultaneous evaluation of the effects of interrelated factors and for exploration of the causes of heterogeneity. It is fairly safe to say that, for the most part, the statistical methods in the current practice of clinical data analysis and reporting are, under a broad definition, the analysis of variance. The development of linear model techniques, after Fisher, greatly facilitates the computations for the analysis of variance.

This book adopts a broad definition for the analysis of variance. It is a mode of analysis in which the data are summarized with the mean and the quality of summarization is measured by the standard error of the mean. Standard error is discussed in detail in Chapter Three. This simple definition streamlines all the technical formalities for the analysis of variance presented in Chapters Four, Five, Six and Seven. Under this broad definition, t-test, chi-square test, multiple regression, logistic regression and survival analysis that are familiar to medical researchers, and nonparametric tests, repeated measures analysis and analysis of variance on complex scales that are not so familiar to medical researchers, become special cases of the analysis of variance. By coming down to the mean and standard error, we are able to fully appreciate the simplicity and flexibility of the analysis of variance technique; by coming down to the mean and standard error, we can clearly see how those mathematically complicated and extravagantly claimed statistical methods lose their ground.

For the blood sugar trial, the result from the analysis of variance are summarized in the following table:

Treatment	Number of patients	Mean	Standard error
Drug D	7	140	30
Placebo	5	102	35

The means and their standard errors are graphically presented with this bar chart:



This example simply shows how the result of analysis of variance is presented with the mean and standard error. The full flavor of the analysis of variance technique is given in Chapters Four and Five.

However, the analysis of variance has serious limitations. Mean is subject to overdue influences from very few extraordinary data values and becomes completely useless when the data distribute in clusters. A potentially more serious limitation comes from the use of standard error. In general, standard error decreases in proportion to the number of observations. When the number of observations is large, the standard error may underestimate the diversity of the data, creating a false impression on the quality of summarization. Finally, in the analysis of variance, a standard error is not specific for the mean. The standard error of a mean is based on the common numerator, called the residual mean sum of squares, which measures the overall quality of summarization with all the means in the analysis. When the primary interest is individual means, the standard errors from the analysis of variance may not be an adequate measure of their precision.

1.3.4 The analysis of dispersion

Patients always respond to therapeutic intervention differently, presumably due to the effects of the uncontrolled factors. When evaluating the effects of a therapeutic intervention, it is always desirable to see what patients responded and what patient did not. Also, because clinical study is, after all, an exploration of the unknown world, it is not surprising that some factors, previously unknown and therefore not controlled in the study, can have significant impact on the patients' response to treatment. Our learning experience comes from identification of those factors and characterization of their effects. The analysis of dispersion serves for these purposes.

Dispersion is the spread of data, and it is a measure of variation. A simple measure of dispersion is the absolute deviations from the mean, where

$$\text{absolute deviation} = |\ \text{data values} - \text{mean}\ |.$$

Absolute deviations closely associate with residuals:

$$\text{residual} = \text{data values} - \text{average}.$$

For both absolute deviation and residual, the mean is the reference point at the center.

In general, dispersion represents the effects of the uncontrolled factors. However, dispersion can be an important measure of treatment effects. For instance, if two groups, comparable at the baseline, demonstrate drastic difference in dispersion after receiving different treatments, the difference in dispersion should be interpreted as the effect of treatment, and the cause of this difference can only be the joint effects of treatment and the uncontrolled factors.

The following table lists the absolute deviations for the blood sugar trial:

Treatment	Absolute deviations							Mean	Std
Drug D	74	18	181	91	52	32	99	78	50
Placebo	14	23	37	5	7			17	12

These absolute deviation values may be viewed as a new set of data, to which any data analytic methods may be applied. Graphical analysis confirms that the deviations clearly distribute in two separate clusters:

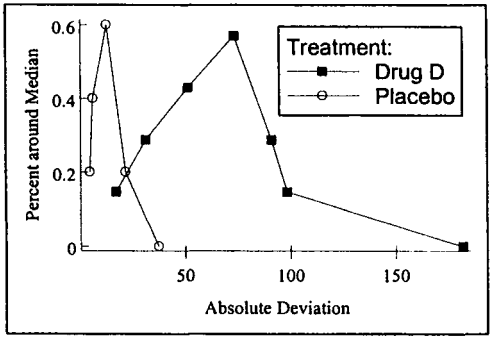


Figure 1.2 Frequency distribution of absolute deviations

Permutation test, which will be discussed in the next section, shows that the deviations in the placebo group are very much different from vast majority of the permutations, with the p-value being 0.008. The means and standard errors from the analysis of variance are 78 and 16 for the drug D group, and 17 and 19 for the placebo group. These analyses converge to a conclusion that the two groups are indeed different in dispersion.

The above analysis of dispersion indicates that the two groups are different. The interpretation of this difference needs to be made in the context of overall study evaluation. If the two groups are comparable with respect to all identifiable factors in the course of study, this difference is dispersion suggests that patients had responded to drug D but quite heterogeneously. If so, the most informative analysis next is the characterization of responders and non-responders.

1.3.5 Permutation test

The idea of permutation was introduced by Ronald A. Fisher to illustrate the concept of fiducial inference in his book entitled *The*

Design of Experiments. In my opinion, permutation test is not effective for clinical data analysis, and it will not be further pursued in this book beyond this section. Permutation test is discussed here mainly because extravagant claims, such as its exactness for small sample sizes, have been associated with this test in the statistical literature. Those claims often confuse clinical researchers and some statisticians as well.

Permutation is a systemic exchange of data values among groups. The following table lists some permutations between the drug D and placebo groups:

Table 1.2 Four Permutations from the Blood Sugar Data

Perm.	*Data permuted to placebo group*					*Sum*
1	89	80	140	108	96	513
2	89	80	140	108	67	484
3	89	80	140	67	123	499
4	89	80	67	123	322	681

Only the data values exchanged to the placebo group are shown; the data values to the drug D group are complementary. The first permutation is the original observations in the placebo group. For the rest of the permutations, the original observations in the drug D group sequentially replace those in the placebo group. The total number of unique permutations is $792 = (12\times11\times10\times9\times8) \div (5\times4\times3\times2\times1)$, which is the number of unique combinations of five out of the twelve blood sugar values.

A permutation test is comparison of permutations. Each permutation is first represented with a single summary measure, for example, the sum of the permutation, and then the permutations are compared through the distribution of their summary measures. Figure 1.3 is the histogram presenting the distribution of 792 sums, each representing the corresponding permutation of the blood sugar data. This distribution, called permutation distribution, is used to count the number of permutations that give rise to extreme summary measures. The reference point of extremity is the summary measure of the original observations. For the blood sugar trial data, the sum of the original data values in the placebo group is 513. 206, 26% of 792, permutations give rise to sums less than 513, and the sums of the remaining permutations

are greater than 513. Therefore, 26% of the permutations are extreme compared to the original data values in the placebo group. The percentage of extreme permutations is referred to as the p-value of the permutation test.

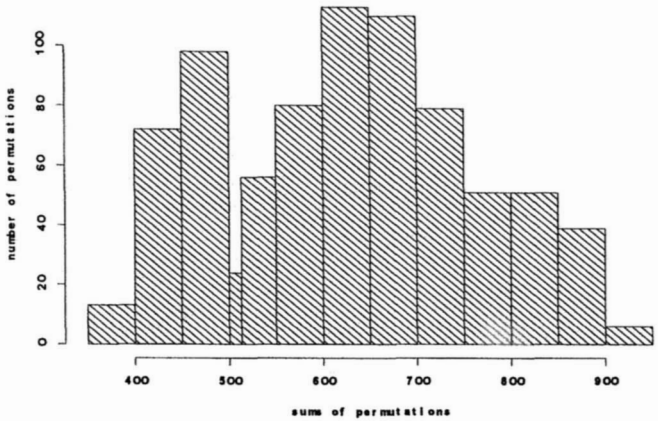


Figure 1.3 Frequency distribution of the sum of permutations

If drug D and placebo are identical, the permutations of their data should also be identical. However, due to the effects of the uncontrolled factors, the actual permutations differ even though drug D and placebo are identical. In the ideal situation where drug D and placebo are identical and the uncontrolled factors are evenly distributed between the two groups, it would be reasonable to expect that the permutations fluctuate around the original observations, and the p-value of the permutation test be 50%. Therefore, the p-value of a permutation test that is significantly smaller than 50% indicates that majority of the permutations are quite different from the original observations. In Fisher's fiducial argument, this small p-value may be used to contradict the hypothesis of equivalent treatments.

Permutation test has limitations. The test relies more on the inter-relationship among the data values than their actual magnitudes, and consequently, important information in the data could be overlooked. A

permutation test may present a small p-value indicating a significant difference between virtually identical groups of data. For instance, the p-value of the permutation test for comparing (1.01, 1.001, 1.001) to (0.99, 0.999, 1.00) is as small as 0.05. Much as it may overstate a trivial difference, a permutation test may underestimate a significant difference. For instance, the p-values of the permutation tests for comparing (32, 54) to (89, 98, 123) and (32, 54) to (1233, 2333, 3344) are the same of 0.10, even though the difference between the second pair of groups is much larger than that of the first pair. Generally speaking, when data distribute in separate clusters, permutation test is not quite useful.

The burden of computation for permutation test can be formidable. In general, permutation tests are technically manageable only when the number of observations is small. However, it is by no means that permutation test is superior to any other statistical methods in presenting underlying information even when the number of available observations is small. Insufficient quantity of observations equally disqualifies the result of any analysis. At most, permutation test is a unique mode of data evaluation that makes it *not* directly comparable to others. The p-value of a permutation test is simply a measure that is meaningful only in the context that test. Neither is this p-value comparable to nor is it more exact than the p-values out of any other valid statistical methods.

This Page Intentionally Left Blank

2
Graphical Data Analysis

Summary

Graphical data analysis must be based on the visualization of individual data values. As opposed to graphical display of summary measures that emphasizes magnitudes, the technical basis of graphical data analysis is simultaneous display of both magnitudes and frequencies of individual data values in order to characterize data distribution. Cross display of multiple graphs by the factors under comparison affords excellent visual contrast for comparison of data distributions. Bar charts are the graph of choice for categorical data. For continuous data, several graphical techniques are available. Picket fence plot and histogram are the graph of choice for direct display of frequencies, box plots show main body and outliers, and cumulative or symmetric cumulative frequency plots are good for comparing multiple distributions on a single graph. Delta plots are the graph of choice for showing changes from baseline. Tree plots are a variant of delta plots and are effective for efficacy analysis. For profile analysis, scatter plots and smoothing techniques are helpful.

2.1 Cross display of data distribution

Graphical data analysis must be based on the visualization of individual data values. A pictorial presentation of individual data values reveals global pictures without losing sight of the diversity of individuals. The technical basis of graphical data analysis is simultaneous visualization of the magnitudes and frequencies of individual data values. In general, these two dimensions fully characterize the distribution of data values.

Graphical data analysis must be distinguished from graphical display of summary measures, such as a bar chart of the means and standard deviations. Graphical display of summary measures emphasizes magnitudes and lacks information on the frequencies of individual data values. Although visualization of summary measures expedites comparison, it does not add any more information to the summary measures themselves. Because no single numerical summary measure is flawless, important information in the data could be overlooked by exclusive use of summary measures. It cannot be overemphasized, therefore, that graphical data analysis must present individual data values.

A full characterization of data distribution is not enough. Graphical data analysis must allow for comparison of multiple data distributions. Just like cross tabulation of summary measures facilitates comparison, a simple and effective graphical analysis technique is cross display of multiple data distributions by the factors under comparison. This organized presentation of multiple data distributions, each characterized by a single graph, provides excellent visual contrasts. For a typical clinical study where the patients are randomly assigned to few treatment groups and followed at a series of clinic visits, the responses to treatment at each visit may be presented with a graph, and then an array of graphs for the entire study may be cross-displayed by treatment and visit:

	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
Treatment 1	Graph –11	Graph –21	Graph –31	Graph –41	Graph –51
Treatment 2	Graph –12	Graph –22	Graph –32	Graph –42	Graph –52
Treatment 3	Graph –13	Graph –23	Graph –33	Graph –43	Graph –53

Compared to cross tabulation of summary measures, the only difference is that summary measures are substituted for graphs.

Thanks to CrossGraphs, an excellent graphical data analysis computer software package from Belmont Research, graphical data analysis by cross display of multiple data distributions has become an easy task. In fact, almost all the graphical analyses in this book are carried out with CrossGraphs.

2.2 Categorical data and continuous data

Categorical data are symbols to denote categories or groups. For example, gender may be recorded with 0 for female and 1 for male. The fact that 1 is greater than 0 has no meaning when these two numbers are chosen to represent gender. What matters is that 1 and 0 are different. Continuous data are numeric values to represent magnitude, usually taking any admissible values within a range. Between categorical and continuous data is grading data, which denote the order or rank of a list of categories or groups.

Once appropriately coded, there is no need to distinguish these three types of data for summarization with numerical measures. It is necessary to make a distinction, however, for graphical data analysis. Categorical data usually have few distinct values each of which may appear frequently, while continuous data have many distinct values, each of which may appear just once. Thus, direct display of the frequencies of categorical data is generally visible and highly informative, whereas direct display of the frequencies of continuous data usually shows little variation from one and is virtually useless. In general, more sophisticated graphical techniques are required to visualize continuous data. The presentation of grading data depends upon the number of categories or groups. When the number is small, the data may be viewed as categorical data; when the number is large, the data may be presented as if they were continuous data.

2.3 Direct display of frequencies

Bar chart, picket fence plot and histogram are demonstrated to display the frequency distribution of data. If appropriately constructed, these graphs authentically represent the information in the data. The drawback is that these graphs tend to be bulky, and it is sometimes difficult to place closely for visual comparison. Bar chart and histogram are perhaps the

most frequently used graphical analysis technique and is familiar to most researchers.

2.3.1 Bar chart for categorical data

Bar chart is the graph of choice for direct display of the distribution of categorical data. The following graph shows the distribution of symptom scores from a trial where the intensity of the symptoms was graded from 0 for none to 3 for severe by an increment of 0.5.

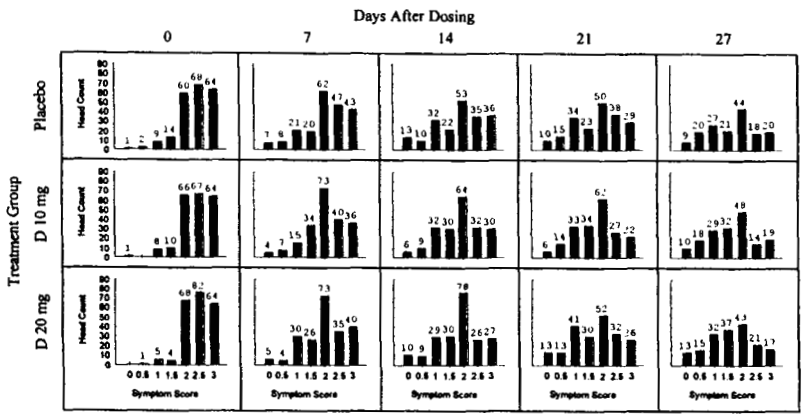


Figure 2.1 Frequency distribution of symptom scores

On the top of each bar is the number of patients in the score category in which the bar stands. Because the number of patients is different among treatment groups and across days, the absolute head counts are not directly comparable. A more appropriate measure for direct comparison is the percentage of patients in each score category, shown in Figure 2.2 on the next page. In that chart, the absolute head counts are still shown on the top of each bar, but the bars are scaled in accordance with the percentage of patients in each category at each of the treatment-day combinations. It is perceivable from these charts that although symptom scores improved in the course of the study as more patients moved down to low score categories, the drug at neither dose demonstrated apparent therapeutic advantage over placebo.

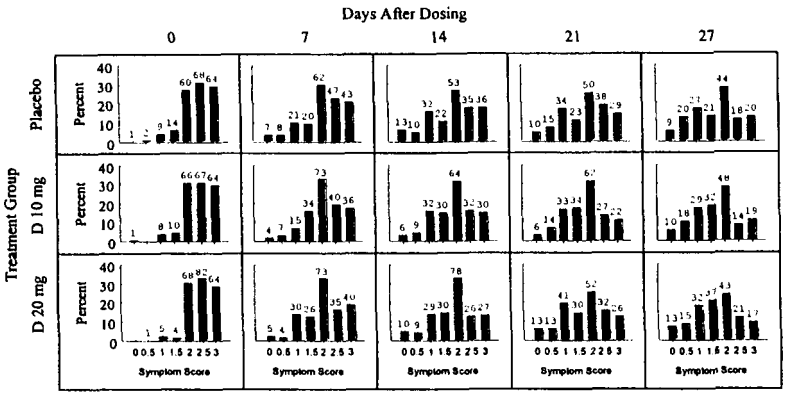


Figure 2.2 Adjusted frequency distribution of symptom scores in percentage

2.3.2 Picket fence plot for continuous data

A picket is a vertical line whose length represents the magnitude of a data value from a patient. A picket fence plot is a list of pickets in descending order. The following graph is a picket fence plot for a single group of data values.

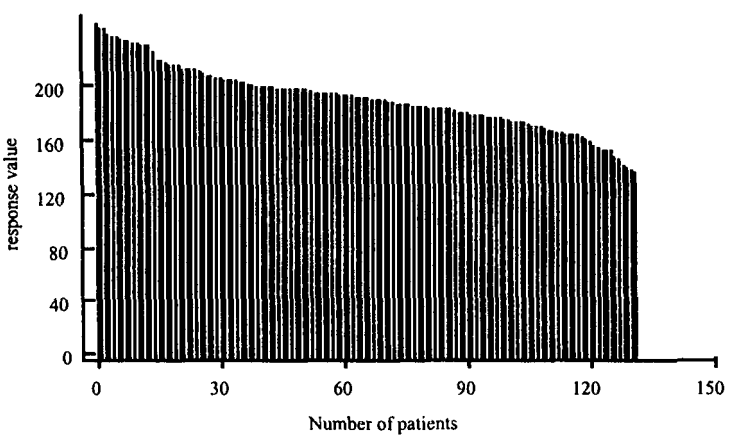


Figure 2.3 Distribution of data values

The y axis marks the data values, and the x axis marks the number of pickets or patients. Like cumulative frequency plot to be discussed in the next section, the scale of x axis may be changed to percentage:

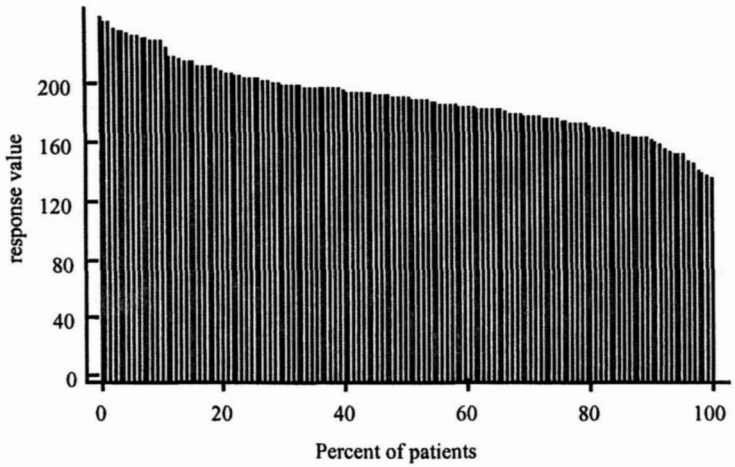


Figure 2.4 Percent distribution of data values

A problem with picket fence plot is that multiple groups cannot be distinctively displayed on a single plot, and this causes inconvenience for visual comparison. For the three groups of data shown in the following plots,

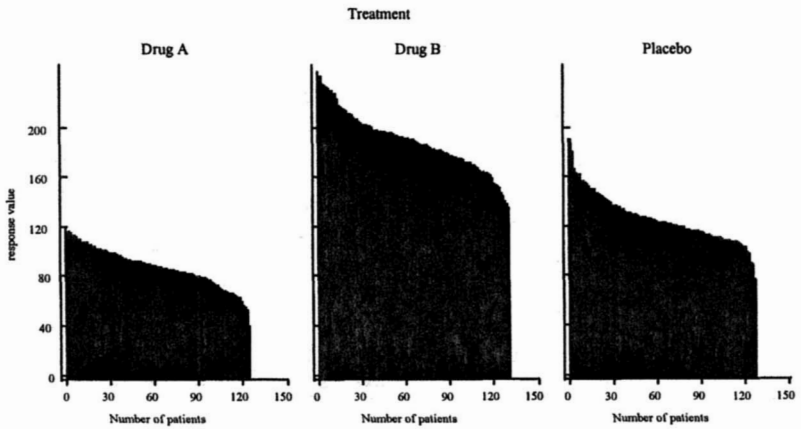


Figure 2.5 Distribution of data values by treatment

for instance, only two groups can be clearly visualized if they are placed in a single plot:

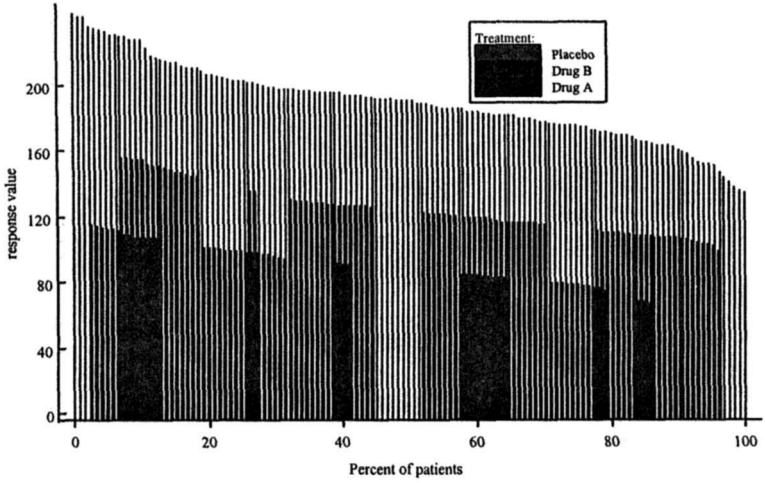


Figure 2.6 Overlay display of three data distributions

A solution to this problem is to display only the top of each picket, instead of the whole line. For the same three groups of data, the picket top presentation clearly depicts the distribution of data:

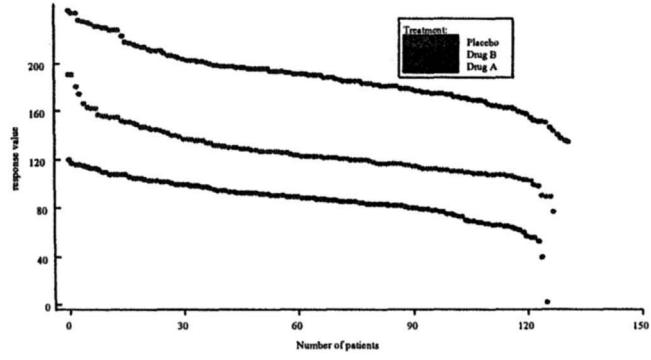
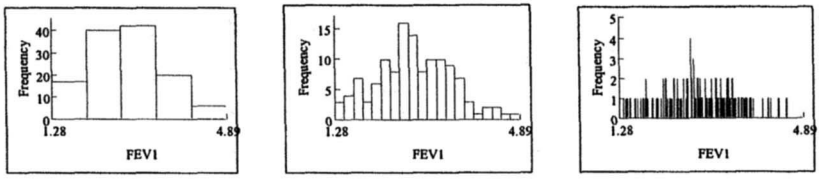


Figure 2.7 Picket top display of three data distributions

2.3.3 Histogram for continuous data

The histogram is also effective to directly display the frequency distribution of continuous data. Continuous data need to be cut into categories in order to be displayed with histogram. Once categories are defined, the construction and appearance of histogram is not different from that of bar chart, both being a direct display of the number or percentage of observations in every category.

The cutting of continuous data into categories affects the appearance of histogram. In general, the greater the number of categories, the smaller the frequency of each category. If the number of categories continues to increase, the histogram will eventually turn into a useless horizontal line. The following histograms are generated from the same data cut evenly into 5, 20, and 500 categories:



Note the scale of frequencies declines as the number of categories increases. There are no hard-and-fast rules for the optimal number of categories to cut continuous data to. The best number has to be determined on a trial-and-error basis. The goal is to maximally present the variation of the data.

2.4 Cumulative frequency plot for continuous data

The cumulative frequency to a value is defined as the percentage of observations *equal to or less than* that value. For instance, the cumulative frequency to 3 in data, (1, 2, 2.5, 3, 4.5, 6), is 4/6, where 4 is the number of data values equal to or less than 3, and 6 is the total number of data values. A cumulative frequency plot is a scatter plot of cumulative frequencies against the corresponding data values:

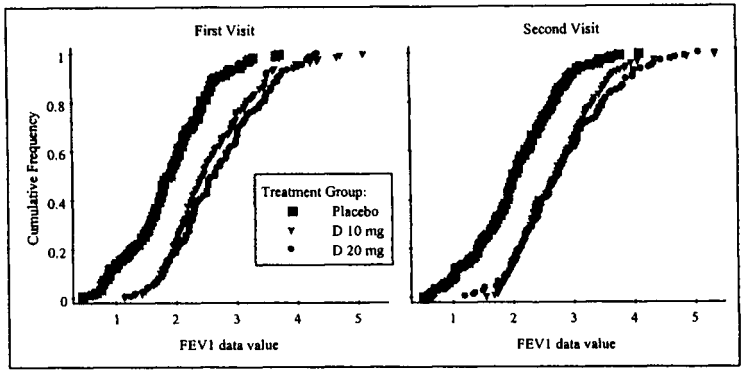


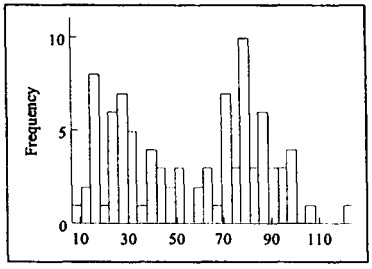
Figure 2.8 Cumulative frequency distribution of FEV1 data by treatment

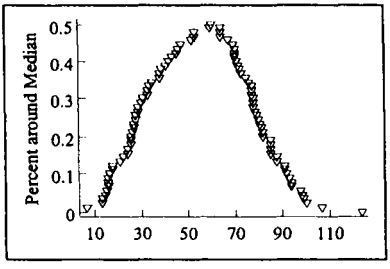
Cumulative frequency curves are non-decreasing. The frequencies continue to increase from zero to one as the value on the horizontal axis increases. In viewing cumulative graphs, attention should be paid to the shape and shift of the curves. In this graph, a shift to the right indicates improvement in FEV1.

A variant of cumulative frequency plot is symmetric cumulative frequency plot, which is a scatter plot of cumulative frequencies against data values equal to or less than the median and complimentary cumulative frequencies against data values greater than the median. A complimentary cumulative frequency of a value is the percentage of observations *greater than* that value. If 60% of data values are equal to or less than 5, for instance, the cumulative frequency of 5 is 60% and the complimentary cumulative frequency is 40% = 1 – 60%, meaning that 40% of data values are greater than 5. The plots in Figure 2.9 on the next page are generated from the same FEV1 data shown in the previous cumulative frequency plots. Symmetric cumulative frequency plot is simply a fold of the corresponding cumulative frequency plot around the median. The folding helps highlight the medians and data spans.

Cumulative frequency plot or its symmetric counterpart affords excellent visual contrast for comparing distributions of continuous data. The data in each group are represented by a curve so that multiple curves can be closely placed on a single graph. This degree of closeness is generally difficult to achieve with histogram. However, cumulative or

symmetric cumulative frequency plots are not the graph of choice for direct display of frequencies. Compare these graphs generated from the same data:





A striking feature is that the data distribute in two clusters. While an appropriately constructed histogram shows the mountains and valley, this change of frequencies is not sensitively reflected on the symmetric cumulative frequency plot. When the number of observations is large, drastic frequency variations may cause only subtle changes in curving. A steep segment corresponds to a cluster of high frequencies, and a flat segment corresponds to a cluster of low frequencies. This lack of sensitivity to change of frequencies could falsely impress a quick browser that the data distribute in a bell shape.

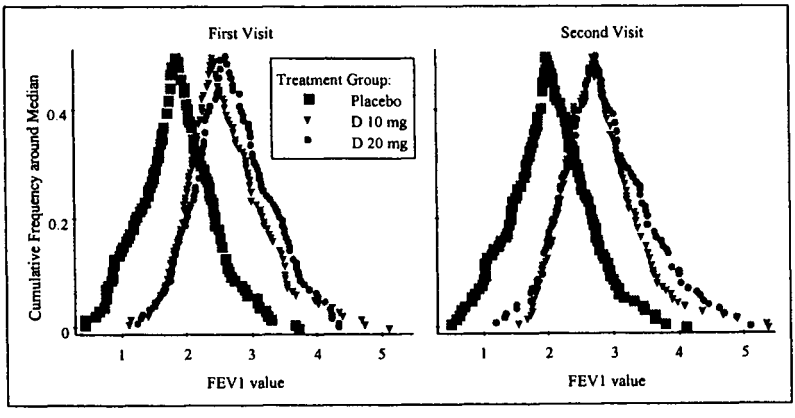
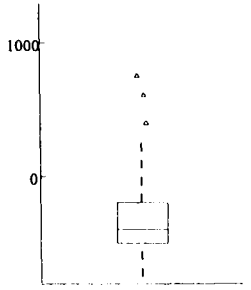


Figure 2.9 Symmetric cumulative frequency distribution of FEV1 data by treatment

2.5 Box plot for showing main body and outliers



Box plot shows the main body of continuous data and individual data values outlying the main body. The box is where the main body resides, covering 50% of data values around the median. The top of the box marks the third quartile (Q3 = 75 percentile), and the bottom marks the first quartile (Q1 = 25 percentile). The horizontal line inside the box represents the median (50 percentile). The vertical dash lines extending from the box are called whiskers. There are several definitions of the maximal whisker length. This book adopts the conventionally definition of 1.5 × (Q3 − Q1). The actual length of a whisker depends upon the range of the data. The whiskers mark the outskirts of the main body. The outskirts vary depending upon the spread of the main body. The dots beyond the whiskers represent outlying data values.

Box plot is a shorthanded description of data distribution. The following box plots are generated from the same FEV1 data previously shown with cumulative frequency plots:

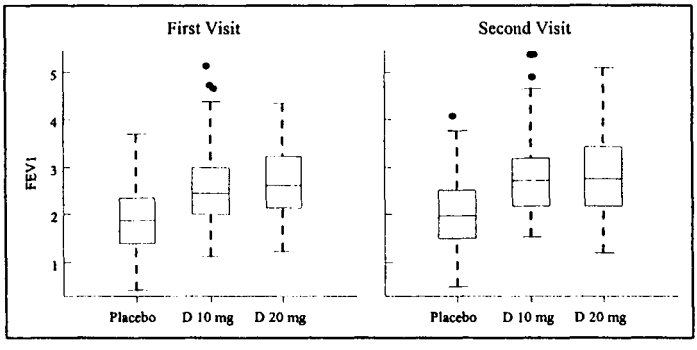


Figure 2.10 Box plots of FEV1 data by treatment and visit

Compared to histogram and cumulative frequency plot, box plot lacks detailed information on the magnitudes and frequencies of data values.

The gathering and spread of data are described with the few percentiles. An advantage is that multiple box plots can be arranged on a single graph, which greatly enhances visual contrast for comparison of data distributions. In addition, the outliers identified on box plots draw attention to rule out possible data errors, investigate the causes of outlying, and study their influence on summary measures.

Another use of box plots is to evaluate the adequacy of summary measures in the background of data distribution. In the following graph, the means and their standard errors are evaluated. The green zones represent the range of mean ± standard error.

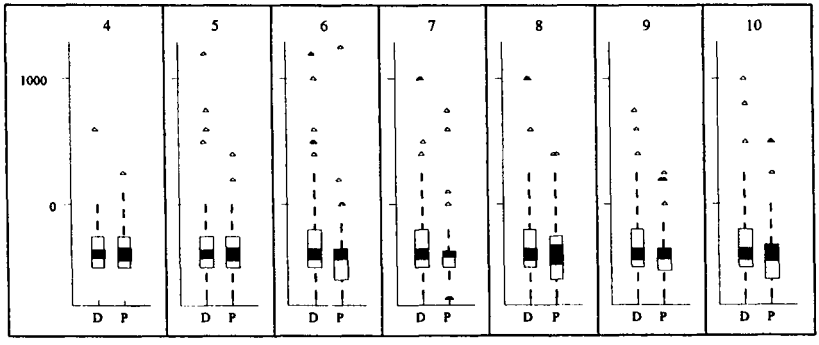


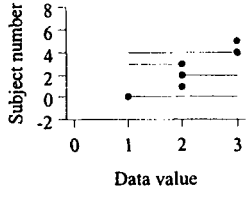
Figure 2.11 Box plots of data distributions, means, and outliers by treatment and visit

The means appear to be similar between D and P across all visits. However, at visits 6, 7, 9 and 10, the range determined by the mean and its standard error does not seem to reside within the main body of the data, and this is especially true for the group on treatment P where more observations are below the mean. This discrepancy demonstrated in these box plots, once again, underscores that important principle that graphical data analysis must be based on the visualization of individual data values, not summary measures. For this example, had the data been summarized only with the means and their standard errors, the true underlying information in the data would have been largely missed.

2.6 Delta plot and tree plot for within-patient contrasts

Delta plot and tree plot are a fascinating graphical technique for presenting data from clinical studies. They are probably the only effective graphical technique to date to present within-patient contrasts. They can be used to demonstrate change from baseline, the phenomenon of regression to the mean and the effect of time on disease fluctuation. Tree plot is particular useful to demonstrate efficacy because it presents the whole spectrum of responses from individual patients.

2.6.1 Delta plot for change from baseline



A delta plot depicts individual subjects each with a horizontal line. Each line has a start value and an end value with the length of the line representing the change from the start to end values. The lines may be sorted by the start values, end values, or changes. In this particular plot, the lines are stacked up in ascending order by the start values.

Figure 2.12 on the next page is the delta plots presenting the change from baseline values of FEV1. The lines are sorted in ascending order by the baseline FEV1 values marked by dots. Lines pointing to the left represent decreases in FEV1, and lines pointing to the right represent increases in FEV1. It is interesting that more improvements in FEV1 occurred in patients with low baseline FEV1 values and more declines in FEV1 occurred in patients with high baseline FEV1 values. This is often referred to as the phenomenon of "regression to the mean." The cause of this phenomenon is likely to be the change of uncontrolled factors in the course of study: The factors that caused high baseline FEV1 measures may have disappeared or changed their status at the time when subsequent FEV1 measures were taken. It is this same phenomenon that cautions physicians not to diagnose a patient as having hypertension until three consecutive assessments weeks apart demonstrate a consistently high blood pressure. In clinical studies, regression to the mean generally reflects the natural fluctuation of the disease process under study, not the effect of therapeutic intervention.

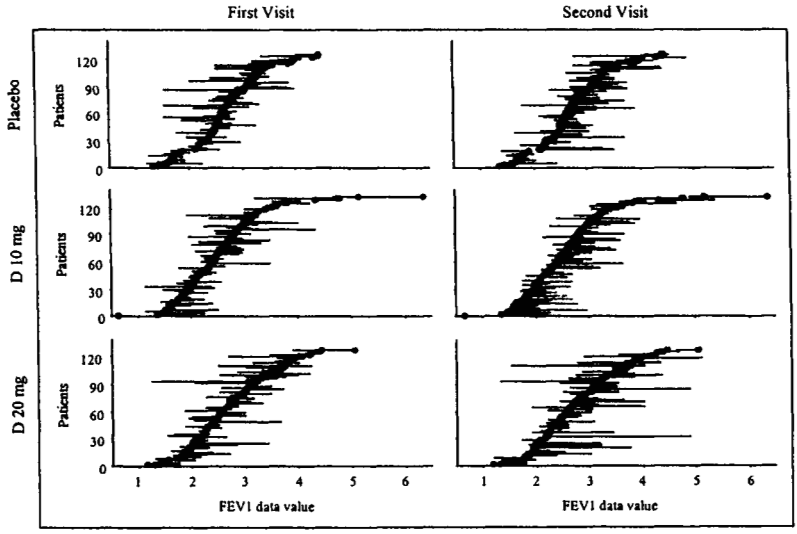


Figure 2.12 Delta plot of FEV1 change from baseline data by baseline, visit and treatment

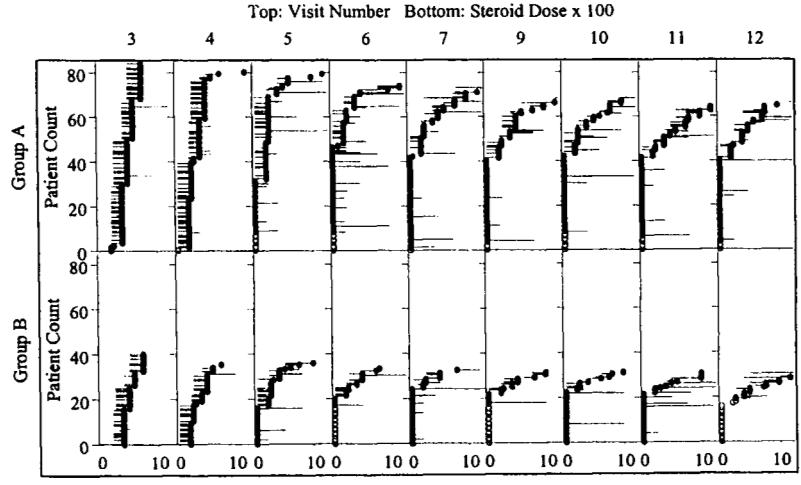


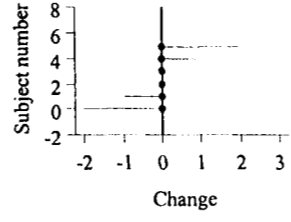
Figure 2.13 Delta plot of change from last visit by visit and treatment

If a baseline measure represents the severity of the illness to treat, delta plot is fascinating in showing the magnitude of response to the therapeutic intervention over the spectrum of severity as measured by the baseline measure. It is not uncommon that while regression to the mean inundates the magnitude of response to ineffective treatment, as demonstrated in Figure 2.12, effective treatment generally results in responses in a single direction, either increase or decrease from the baseline, not a mixture of both as in Figure 2.12. Regression to the mean, however, still plays its role in the background in that the magnitude of responses spreads out over a spectrum. In general, effective treatments work better in sicker patients. In other words, the magnitude of responses from critically ill patients tends to be greater than that from patients whose illness is moderately severe.

When patients are followed for a period of time, delta plots can be very insightful by presenting the sequential changes of response in the time course. Figure 2.13 on the previous page demonstrates the dose changes of corticosteroids from a steroid-sparing study, in which steroids were sequentially adjusted to maintain symptom control within a specified range. Instead of a fixed baseline for all visits, the baseline measures at each visit are the steroid doses at last visit, depicted by dots. Lines pointing to the left depict decreases in steroid doses, and lines pointing to the right depict increases in steroid doses. It is clearly shown that the steroid doses were reduced for most patients at the first 2 or 3 visits and started to fluctuate. The fluctuation may be largely attributed to the effects of the uncontrolled factors and reflect the waxing and weaning nature of the disease. Overall, there is no apparent difference between groups A and B during the entire course of study.

2.6.2 Tree plot for efficacy analysis

Tree plot is a variant of delta plot, where the start values are aligned vertically at zero, and the lines are sorted in ascending order by changes. Tree plots are excellent for demonstrating efficacy by showing the number of patients who do and do not



benefit from treatment and the extent of beneficial and non-beneficial effects.

Figure 2.14 is tree plots showing change from baseline measures of FEV1 by visit and treatment. The blue lines away from zero to the right represent the patients with improving FEV1 from the baseline, the red lines away from zero to the left represent the patients with deteriorating FEV1 from the baseline, and the hinge represents the patients with no change in FEV1 from the baseline. These tree plots give a complete list of individual patients' responses, whereas the mean responses are only the net result of those positive and negative responses.

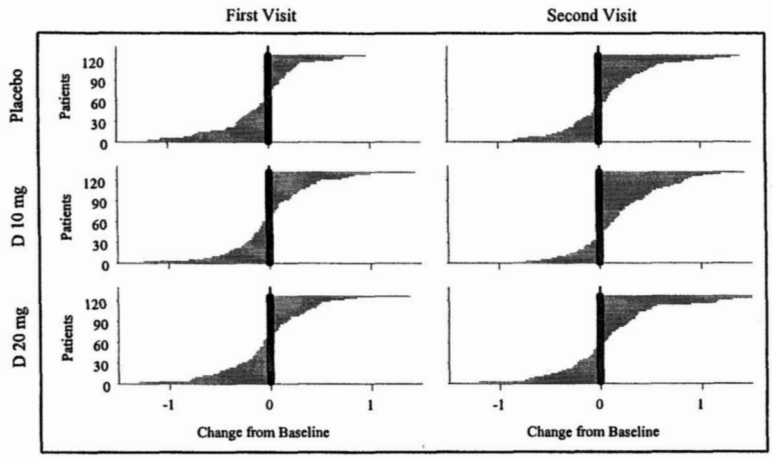


Figure 2.14 Tree plot of change from baseline by visit and treatment

Tree plots are also useful to show efficacy for studies where each patient receives multiple treatments. A 2 x 2 cross-over study is an example, where a patient receives two treatments each in a period. Figure 2.15 on the next page is a tree plot that presents the result from a 2 x 2 cross-over study where each patient receives both placebo and treatment. In this tree graph, the start values aligned vertically at zero are the responses in the period on placebo, and the end values are the responses in the period on treatment; the length between each pair of start and end values represents the within-patient contrast between the periods on

treatment and placebo. The color of the lines indicates other characteristics of the patients. The overall treatment effects are the sum of all the negative and positive contrasts.

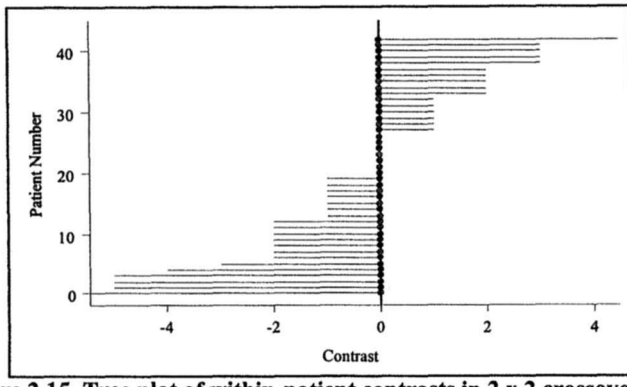


Figure 2.15 Tree plot of within-patient contrasts in 2 x 2 crossover trial

2.7 Scatter plot for profile analysis

Most clinical studies are longitudinal where the patients are assigned to different treatment groups and followed over a time course. For these studies, it is much more informative and reliable to evaluate the response profiles over the course of the study than to focus merely on the responses at few static time points. A simple and effective presentation of response profiles is to draw a response curve for each patient on a scatter plot. In the scatter plots shown in Figure 2.16, each patient's steroid doses at all attended visits are linked with straight lines. When the number of patients is small as in the placebo group, this method is effective. However, when the number of patients is as large as it is in the treatment group, it becomes difficult to visually discern any pattern out of this chaotic weave of lines. In this situation, smoothing techniques are helpful.

In essence, smoothing is local averaging. The whole range of the data values is cut into several intervals, and the data in each interval are represented by their average. A smooth line is then drawn connecting the averages of these intervals. By smoothing, we sacrifice information in exchange for visual clarity. How much information we would like to preserve depends upon the number of intervals that the data are cut into. The greater the number of intervals, the closer the smoothed curve is to the

observed data. With the same data displayed in the Figure 2.16, the scatter plots in Figure 2.17 present the smoothed profiles.

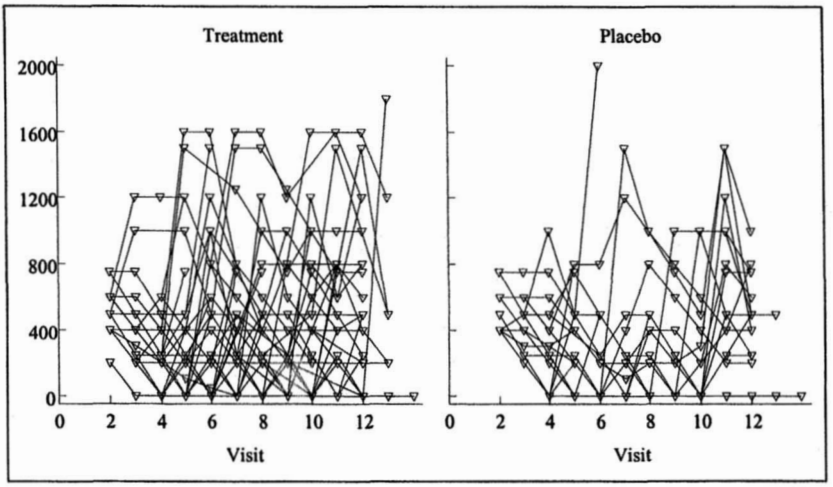


Figure 2.16 Individual response profiles by direct link of data values over time

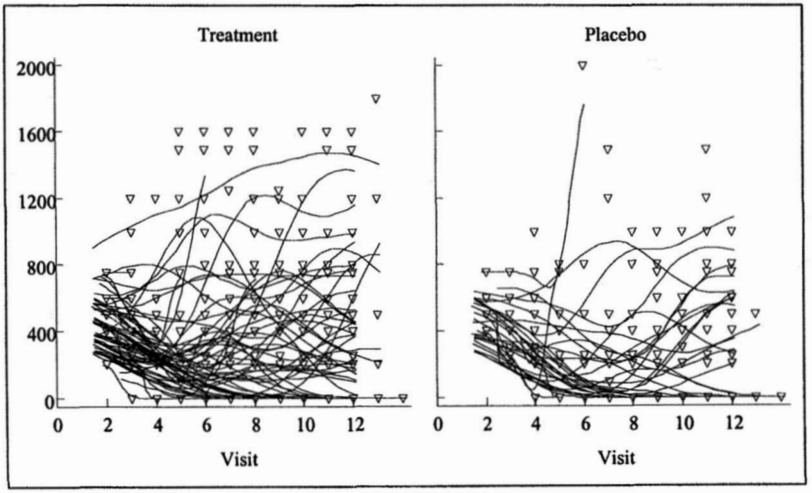


Figure 2.17 Individual response profiles after smoothing over data values over time

By smoothing, some details fall into the background and a clear pattern of response profiles emerges. It is clearly shown that the steroid doses for most patients are reduced before visit 6, and after that, many patients' steroid doses had to be up-titrated. This pattern is shown in both treatment and placebo groups, and it is difficult to appreciate any difference between the two groups because the numbers of patients are different.

This Page Intentionally Left Blank

3
Data Analysis with Summary Measures

Summary

Cross tabulation and display of summary measures by the factors under study expedite comparison. Bar charts are good for showing magnitudes, and line-scatter plots are good for showing trends. Commonly used summary measures are the number of observations, mean, median, standard deviation, average deviation, and standard error. Except for categorical data in a single category where the mean and number of observations are sufficient to characterize the data, all summary measures have limitations that prevent them from capturing the complete information in the data. The mean is subject to overdue influences from few outlying data values, and neither the mean nor median is apt to summarize data that distribute in clusters. The standard error has no simple interpretation, and if it is used to represent the effects of uncontrolled causes, the standard error can be misleading when the number of observations is large.

3.1 Cross tabulation and display of summary measures

Cross tabulation of summary measures presents summary measures in a table by the factors under comparison. The following table is a summary

of the FEV1 data from a study:

Table 3.1 Summary of FEV1 Data

	First Visit			Second Visit		
Measure	**Placebo**	**D 10 mg**	**D 20 mg**	**Placebo**	**D 10mg**	**D 20mg**
N	126	132	128	126	132	128
Mean	2.6547	2.5739	2.7060	2.8253	2.7699	2.8430
Std	0.6919	0.7556	0.7524	0.7376	0.7353	0.827+2
Std/Mean	26%	29%	28%	26%	27%	29%
5th Percentile	1.54	1.57	1.61	1.6	1.78	1.71
25th Percentile	2.21	2.03	2.16	2.32	2.20	2.20
Median	2.665	2.44	2.64	2.78	2.715	2.76
75th Percentile	3.15	3.00	3.22	3.34	3.21	3.43
95th Percentile	3.87	4.01	4.12	4.09	3.98	4.37

This table is designed for a quick browse of the summary measures across treatment groups at each visit.

Three panels of summary measures are tabulated. The first is the number of observations, denoted by N. This measure of study size is most critical in determining the strength of observed evidence. Without sufficient observations, nothing else matters. The second panel consists of three measures. They are the mean, standard deviation, and their ratio. The mean is a fair representation of the contributions from individual data values. The standard deviation measures the average deviation of the mean from the data. The ratio of the mean and standard deviation is an assessment of the quality of summarization with the mean. The third panel is a collect of percentiles for a scratch of the data distributions.

Human eyes may handle a 2 x 2 table easily, but they soon become powerless for mastering any table larger than 4 x 4. When a large number of summary measures are compared across multiple categories, graphical presentation is no longer an option but a necessity. The graphical technique of choice is cross display of summary measures by the factors under comparison.

Bar charts are convenient for a clear display of the magnitudes of summary measures, especially when they are numerically close. The following vertical bar charts show the means, medians and standard deviations in the previous table:

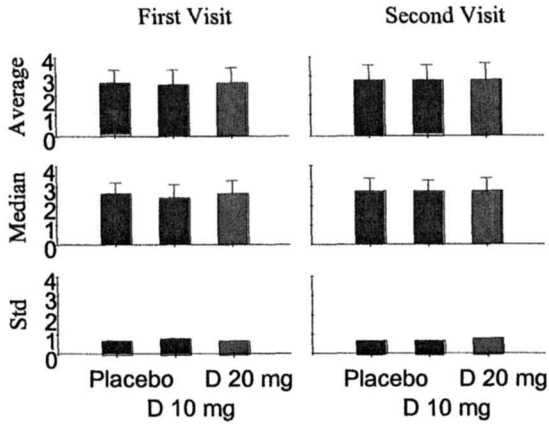


Figure 3.1 Bar chart display of summary measures

The vertical lines that extend from the top of the bars represent the standard deviations of the means in the first row and the average deviations of the medians in the second row. The average deviation will be defined in section 3.4. When the primary interest is looking for trends, summary measures may be shown with a line-scatter plot. Figure 3.2 on the next page is a line-scatter plot that depicts the means and their standard errors in each treatment group over the time course of the study. Each dot represents a mean value, and the vertical bars extending from the dot in both directions represent the standard deviation of the mean. This plot allows the viewer to compare the mean responses over time and across groups. The mean responses in all four groups improved over time. The placebo group has the worst mean responses across all time points, and the best mean responses are shown in the group on competitor's medicine. The mean responses to drug D at both doses are similar, and their effects lie between the effects of placebo and competitor's medicine.

It is concise to characterize groups of data with few summary measures, and communication with summary measures is precise. Researchers should be aware, however, that every summary measure has limitations that prevent it from capturing all the information in the data.

Exclusive use of summary measures in data analysis can be misleading. A good practice is to visualize data with graphical techniques before attempting any summary measures. If summary measures are chosen to represent the data, they must not be presented in isolation. Instead, they must be presented together with at least a measure of precision to indicate the quality of summarization.

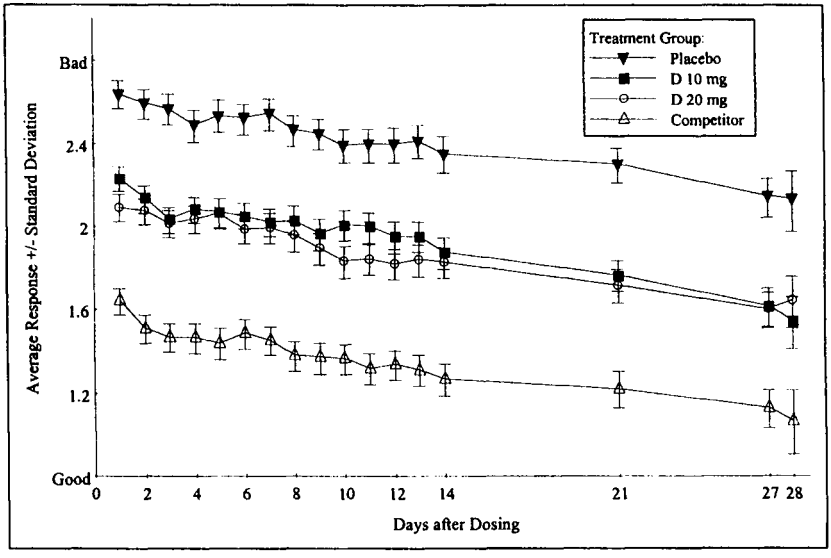


Figure 3.2 Line-scatter plot of means and their standard deviations by time and treatment

3.2 Number of patients for reliability and robustness

The number of patients, also known as study size and sample size, is perhaps the single most important parameter in determining the strength of observed evidence, and it is a direct measurement of the reliability and robustness of study results. Reliability is the repeatability of a result in a series of similar studies. Robustness is the stability of a result against minor fluctuations.

The reliability and robustness of results from small studies are generally poor. Small studies suffer from lack of representability. The patients enrolled in study do not full represent the patient population, and

the result cannot be generalized to guide management of other patients. Also, due to significant differences from sample to sample, small studies often present fragments of the whole picture, and the results are not reproducible. Small studies also suffer from confounding from the uncontrolled factors. Imagine that if an uncontrolled factor exerts some significant effect on two of the ten patients in a study, that effect may significantly alter the study result, to which the contribution from the two patients is as high as 20%. If the effect of that uncontrolled factor is not separable from the effects of treatment, that uncontrolled factor will seriously confound with the effects of treatment.

The shortcomings of small studies are overcome with large studies. The results of large studies are more reproducible due to diminished differences among samples. Intuitively, two 50% samples of 100 patients should be much more alike than two 10% samples. The result of large study is also more robust against the effects of the uncontrolled factors. Much as a drop of dye will not change the color of the Pacific Ocean, the effect of any particular uncontrolled factor on few patients are diluted in proportion to the total number of patients, even though that effect may be significant.

Ideally, the larger the study, the better the result. In reality, clinical study cannot be infinitely large due to the limitation of resource. Unfortunately, a universally accepted standard to determine an adequate study size does not exist. Some criteria for determination of sample size are discussed in Chapter Eight. The key is to strike a balance between variation and certainty. If large variation is expected, a large study may be necessary to demonstrate a pattern with satisfactory degree of certainty so that claim can be made and comfortably accepted.

The determination of sample size with statistical power is a utopia based on the unrealistic statistical theory of Neyman and Pearson. The fantasy and absurdity of their theory and statistical power are exposed in Chapter Ten. Although statistical power has no practical meaning whatsoever, it is the current practice that sample size must be calculated from statistical power and documented in the research protocol. Knowing that officials are in charge even they are necessarily evil, researchers have no choice but comply in order to gain research funding, permission for marketing and publication. For all practical purposes, the entire power

calculation may be viewed as a pipeline of pure mathematical manipulations, which starts off with few presumptuous numbers and ends up with few different numbers. To compute sample size, the presumptuous numbers are power, a standard deviation and a "clinically significant difference." To avoid unnecessary trouble, power must be aimed high, and there is little room to play. 90% to 99% are generally regarded to be high. The trick to get the desired sample size for your practical and scientific purposes is to find the right standard deviation and declare a "clinically significant difference" that is right for you and acceptable by others. While a search in literature should provide sufficient options for a standard deviation, it often requires careful negotiation for a "clinically significant difference." Most of the time, this strategy works out well. First, figure out the target sample size for your study. Then, use a computer to try on different combinations of power, standard deviation and "clinically significant difference" till you find a right combination that produces a sample size close to your target. Unfortunately, as long as statistical testing based on the theory of Neyman and Pearson is required in clinical research by authority, this crying, nevertheless harmless game, will continue on and on.

3.3 Mean and number of observations for categorical data

If positive responses to a category are coded with 1 and 0 otherwise, the mean and number of observations are sufficient to characterize the data distribution. The mean is the frequency or percentage of positive responses. The following table

Patient	A	B	C
1	0	1	0
2	0	1	0
3	1	0	0
4	0	1	0
Mean	0.25	0.75	0

shows that one patient responds to category A, three to B and none to C. The means are exactly the percentages of patients who responded positively to categories A, B, and C.

The number of observations indicates the strength of observed evidence, and for a single category, the mean completely characterizes the distribution of data. For categorical data in a single category, both the average deviation (ad) and standard deviation (std) are completely determined by the mean:

$$ad = 2p(1-p) \text{ and } std = \sqrt{p(1-p)},$$

where p denotes the mean response in the category. Thus, these two deviation measures have no more information than the mean itself for characterization of categorical data in a single category.

3.4 Frequently used summary measures

This section discusses the mean, median, standard deviation, average deviation and standard error. In the current practice, the mean and its standard error are most commonly used in publications. While median enjoys the same popularity as mean, standard deviation is losing favor. Average deviation is seldom used. The conclusion of this section is that the mean and average deviation are better measures than the median and standard deviation, and standard error is a problematic measure without clear interpretation.

3.4.1 Mean and median

The mean, or average, is the most commonly used summary measure defined as:

$$\text{mean} = \frac{\text{sum of data values}}{\text{number of data values}}.$$

Each data value makes a fair contribution to the mean. The relative contribution from each data value decreases as the total number of data values gets large. The mean of a large number of data values is fairly robust against fluctuations caused by the uncontrolled factors. The means are comparable across groups with different numbers of observations, whereas the sums are not. Generally speaking, if data distribute within a narrow range, the mean can be a fairly good representation of the data.

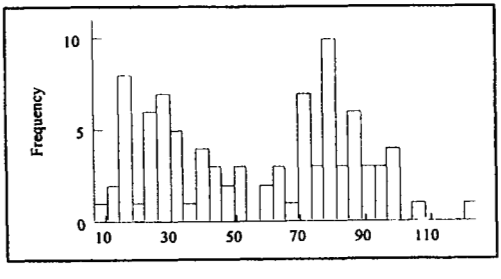
A problem is that the mean is subject to overdue influence from even a single extraordinary value, especially when the number of observations is not very large. For instance, the mean of (1, 2, 3, 4, 100), which is 22, is completely carried away from majority of the data values (1, 2, 3, 4) by the single outlying value of 100. Of course, when the number of data values is large, the mean is relatively robust to a small number of outliers because their contributions to the mean are diluted in proportion to the number of observations.

Comparing to mean, the median is robust to a small number of outliers even though the number of observations is small, making it a candidate to substitute for the mean. But median is much less sensitive a measure than mean. Data contribute to the median by their ranks, not magnitudes. Consider the following data:

Group	Data							Mean	Median
D1	1	5	9	10	11	13	13	8.86	10
D2	2	5	6	10	14	25	30	13.14	10

These two groups of data appear to be quite different. While the message is more or less picked up by the means, it is completely missed by the medians. Although robustness is an attractive property for a summary measure to have, one should carefully balance that property with its low sensitivity and the consequent loss of information.

Both mean and median are poor measures to characterize data that distribute in clusters. For the data shown in the following histogram,



the mean is 57, and the median is 62. Both the mean and median hit the valley and overlook the mountains, and thus, neither truly represents the data.

3.4.2 Standard deviation and average deviation

The standard deviation (std) of a summary measure is defined as

$$std = \sqrt{\frac{\text{sum (data values - summary measure)}^2}{\text{number of data values}}},$$

and the average deviation (ad) of a summary measure is defined as

$$ad = \frac{\text{sum (|data values - summary measure|)}}{\text{number of data values}}.$$

Both the standard deviation and average deviation measure how the summary measure deviates from the observations on average. Although the average deviation appears to be more straightforward a measure, the standard deviation is more commonly used in the current practice of data analysis and reporting.

Both measures can be directly used to indicate the quality of summarization. However, since the magnitude of both standard deviation and average deviation depends upon the magnitude of the summary measure, a more reasonable measure is the ratio

$$\frac{\text{standard deviation}}{\text{summary measure}} \text{ or } \frac{\text{average deviation}}{\text{summary measure}}.$$

A large standard deviation or average deviation suggests two possibilities. One is that the summary measure is not a good measure to capture the information in the data. In that case, another measure may have to be attempted. More often is the other possibility that the data vary too much to be summarized with a single measure. In this situation, the data need to be further analyzed to identify the cause of variation. If the variation is largely due to the uncontrolled factors, critical factors may

have to be identified and controlled in future studies. If the variation is due to treatment, then the variation becomes an important measure for the treatment effects, indicating diverse responses to the treatment.

3.4.3 Standard error

The standard error (stderr) is specific for the mean, defined as

$$stderr = \frac{\text{standard deviation of the mean}}{\sqrt{\text{number of data values}}} .$$

It is a composite measure, recognizing the importance of deviation in assessing the precision of summarization and the number of observations in evaluating the strength of observed evidence. The standard error is a derived measure in mathematical statistics. It associates exclusively with the mean and is often referred to as the variance of the mean. The mathematical operation is stipulated to follow the rule of probability theory. Let $x_1, x_2, ..., x_n$ denote n independent values, and let σ^2 denote the variance of each data value, then var $(\Sigma x_i/n) = \Sigma$ var$(x_i)/n^2 = \sigma^2/n$. Independence and variance to each data value are the stipulation to fit in the theory of probability. This unfortunate attachment of statistics to probability theory is discussed in Chapter Ten.

Unlike standard deviation, the standard error of mean does not have a straightforward interpretation. People who have a touch of reality generally have trouble to understand the concept that a single data value, such as the mean, has any variation. In statistical textbooks, this concept is often explained with an imagined series of studies. If the same study is repeated many times, the variance of the mean implies how those means would vary.

For a single study, the meaning of standard error is obscured. As a measure, it is not nearly as good as standard deviation. For instance, the mean of data consisting of fifty 1's and fifty -1's is 0, and its standard deviation is 1. While the standard deviation accurately reflects the average deviation of the mean from the data, the standard error of the mean, $0.1 = 1/\sqrt{100}$, gives no clue as to how well the mean represents the data. Moreover, for a group of data with five thousand 1's and five thousand -1's, while the mean and standard deviation remain the same, the standard

error now becomes $0.01 = 1/\sqrt{10{,}000}$, which creates a false impression that the mean be a quite good summary of the data.

If standard error is used to represent the effects of the uncontrolled factors, the result of statistical analysis can be seriously misleading. The following symmetric cumulative frequency plots show the distributions of two groups of data with increasing numbers of observations:

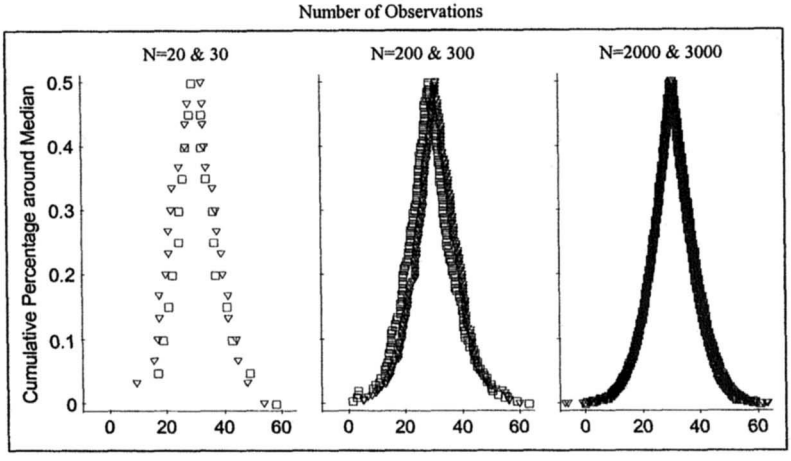


Figure 3.3 Frequency distributions of data in two groups with increasing number of data values

It is quite obvious that the two groups are almost identical, and this is confirmed with increasing number of observations. However, in terms of standard errors, as shown in the following table,

	N = 20/30		N = 200/300		N = 2000/3000	
Group	**Mean**	**Stderr**	**Mean**	**Stderr**	**Mean**	**Stderr**
A	31.40	2.33	28.63	0.75	29.94	0.21
B	30.55	2.02	31.24	0.57	30.98	0.18

the opposite conclusion that the two groups are different may be drawn. The standard error quickly fades away as the number of observations becomes large. Eventually, the magnitude becomes so small that any

trivial difference of the means appears to be significant when compared to that fading standard error.

The standard error is often used to claim inconsequential differences. It tends to be used more often than standard deviation in research presentations because standard error is generally smaller than the corresponding standard deviation. A public secret is that with standard error, any hypothesis of no difference can be rejected in principle by simply increasing the number of observations. This problem with standard error in the analysis of large studies has led to propositions of limiting the number of observations to prevent false claims even though the resource allows for more observations. It has been proposed that the sample size of a study should not be as large as the researcher wishes to be, and it must be determined with statistical power based on the theory of Neyman and Pearson. While deliberately using a poor measure is bad enough, resorting to an unrealistic theory to justify a poor measure is hopeless.

4
The Analysis of Variance

Summary

The analysis of variance summarizes data with the mean, and the quality of summarization is measured with the standard error. The major use is simultaneous evaluation of multiple interrelated factors. The basic operation is grouping and curve fitting. When multiple factors are evaluated simultaneously, any specific effect may be quantified with up to three types of measures, depending upon the relationship among the factors. Type I measures quantify the effects of single factor in isolation from others. More useful are type II and III measures. While type II measures quantify the additive effects, type III measures quantify the joint effects of multiple factors. Both type II and III measures may be combined. The combined measures are referred to as marginal or least squares means. The result of analysis of variance is best presented by graphical display of the means and their standard errors.

4.1 The method

The analysis of variance (ANOVA) summarizes data with the means, and the quality of summarization is quantified with the standard errors. This broad definition of analysis of variance applies to all the analytical

methods to be discussed in Chapters Four, Five and Six, including multiple regression analysis, logistic regression analysis, categorical data analysis, survival data analysis and repeated measures analysis. The definition of mean in the analysis of variance is the same as that in Chapter Three. The standard error is, however, defined quite differently. As opposed to the conventional standard error that is specific to the mean, the standard error in the analysis of variance is based on a global measure for all the means. For a mean in the analysis of variance, its standard error is defined as

$$\text{standard error in anova} = \sqrt{\frac{\text{mean residual sum of squares}}{\text{number of observations the mean represents}}},$$

where the mean residual sum of squares measures the average deviation of ***all the means***, not any specific mean, from the data.

The arithmetic from steps A to D in the following table illustrates the method. The data are from the blood sugar trial, and the purpose is to compare drug D to placebo.

Table 4.1 Arithmetic in the Analysis of Variance

Site	Group	Original data values and squared differences							Summary
A	Drug D	67	123	322	232	89	109	42	Mean=140.57
	Placebo	89	80	140	108	96			Mean=102.60
B	Drug D	140	140	140	140	140	140	140	Grand mean
	Placebo	102	102	102	102	102			=124.75
C	Drug D	250	250	250	250	250	250	250	Sum of squares
	Placebo	490	490	490	490	490			=4205.34
D	Drug D	5412	308	32916	8359	2659	996	9716	Sum of squares
	Placebo	184	510	1398	29	43			=62536.91

Step A is list of the original data values and group means. At step B, the original data are replaced with the group means. Step B is critical, where the group means are chosen to characterize the observations in treatment groups. Had another summary measure been chosen, the result would have been different. The grand mean is the average of all observations regardless of treatment. Tabulated at step C are the squared differences between the group means and grand mean; for instance, $(140 - 124)^2 =$

250. The sum of these squared differences, known as sum of squares, presents an alternative measure of treatment effects. If the effects of drug D and placebo are the same, the sum of squares should be small. Step D lists the squared differences between the original observations and group means; for instance, $(67 - 140)^2 = 5412$ in drug D group and $(89 - 102)^2 = 184$ in placebo group. The sum of these squared differences, known as residual sum of squares, measures the average deviation of group means from the original observations and is generally interpreted as the variation caused by the uncontrolled factors.

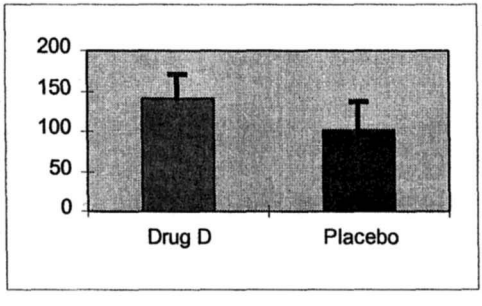
Treatment effects are evaluated by comparing the group means to their standard errors. The standard errors are derived from the residual sum of squares at step D. The first is to compute the mean residual sum of squares:

$$\text{Mean residual sum of squares: } \frac{\text{residual sum of squares}}{\text{degree of freedom}} = \frac{62536.91}{10} = 6253.69 .$$

Then the standard errors of group means are simply

$$\text{Drug D: } \sqrt{\frac{6253.69}{7}} = 30 , \text{ and Placebo: } \sqrt{\frac{6253.69}{5}} = 35 .$$

Comparison can be easily made by visual contrast:



Treatment effects can also be evaluated by comparing the mean sum of squares of treatment to the residual mean sum of squares. The mean sum of squares is sum of squares divided by the number of essential pairwise

contrasts, known as the degree of freedom. Essential pairwise contrasts are mutually exclusive and form a basis for deriving any other pairwise contrasts. Between treatment groups, there is only one essential contrast, 140 – 102, and thus, the degree of freedom for treatment effects is 1. Within treatment groups, there are four essential pairwise contrasts in the placebo group, such as 89 – 80, 89 – 140, 89 – 108, and 89 – 96, and six in the drug D group, such as 67 – 123, 123 – 322, 322 – 232, 232 – 89, 89 – 109, and 109 – 42. Thus, the degree of freedom for residual sum of squares, the effects of the uncontrolled factors, is 10 = 4 + 6. These ten contrasts are essential in the sense that any within-group contrast can be derived from them; for example, in the placebo group, 80 – 140 = (89 – 140) – (89 – 80). Essential contrasts are not unique. In the placebo group, for instance, another set of four essential contrasts may be 89 – 80, 80 – 140, 140 – 108, and 108 – 96. The mean sum of squares is invariant upon the formation of essential contrasts. Any mean sum of squares with more than one degree of freedom is an agglomerate measure. As opposed to specific comparisons of group means, the mean sum of squares measures the average of a class of pairwise comparisons. In terms of mean sum of squares, the result of analysis of variance may be summarized in a table, known as ANOVA table:

ANOVA Table

Source	*DF*	*SS*	*MS*	*F*	*P*
Treatment	1	4205	4205	0.67	0.43
Residual	10	62537	6254		

SS: sum of squares, MS: mean sum of squares,
DF: degree of freedom, F: ratio over residual MS, P: p-value

This table states that the variation due to treatment is not significantly larger than the variation caused by the uncontrolled factors, and therefore, the effects of treatment are not significant.

4.2 Basic operations in the analysis of variance

The analysis of variance allows for simultaneous evaluation of the effects of multiple factors, which is very convenient and effective for investigating heterogeneity. However, the analysis of variance does not automatically attribute effects duly to the causal factor or factors, and the

method itself does not offer any mechanism for the control of confounding. Researchers must be constantly conscious of the fact that the analysis of variance as well as any other statistical analytic methods are techniques for human purposes. Therefore, researchers must know their purposes before data analysis.

Nowadays the analysis of variance can be done automatically with a computer, and what the user has to do is to select factors and plug them into the computer. This wide availability of computer programs for carrying out analysis of variance has spread misconception of the technique and prompted misuse. Confounding may be introduced by unscrupulously adding factors into the analysis. As an extreme example, if a volcano erupts in Los Angeles at the same time when the crystal ball lights up year 2000 in New York City, the analysis of variance can establish a perfect correlation between these two events. Misuse of analysis of variance, like this, might have accounted for many of shocking conclusions that hit the headline news and made the reporting "scientist" a national kangaroo. While the absurdity shown in this example is obvious, the practice of using computer to select factors for the analysis of variance is potentially dangerous because the technical complexity conceals the absurdity of same nature from untrained people. Stepwise multiple regressions and stepwise logistic regressions are often encountered in medical literature for selecting a set of factors to account for the variation of data. These selection algorithms use some arbitrary thresholds on the reduction of mean residual sum of squares to include or eliminate factors. By using these algorithms, data analysis is turned into a game of gambling on those thresholds, not different from looking at crystal ball in strategic planning or tossing coin in decision making. Man-made computers will continue to improve, but they will never take away careful thinking of research purposes from the computer between our ears.

The aim of this section is to demonstrate what we should think for the analysis of variance, what the analysis of variance can do for us, and how the computation in analysis of variance can be carried out with the linear model technique.

4.2.1 Controlled versus uncontrolled factors

It is important to distinguish controlled from uncontrolled factors. The controlled factors determine the assignment of patients to treatment. They are the backbone of study design. By including these factors into analysis, the mechanism for the control of confounding inherited in the design is fully utilized. The uncontrolled factors do not relate to patient assignment. They are called covariates if recorded. In general, the uncontrolled factors confound with the controlled factors. The only effective means to control the confounding is equal distribution of the uncontrolled factors among treatment groups. In general, the magnitude of confounding from the uncontrolled factors cannot be altered unless they are controlled by stratification. Only occasionally, incorporating covariates into analysis may reduce the magnitude of confounding.

The following table illustrates the structure of a ten-center trial:

Table 4.2 Structure of a Multicenter Trial

Factors that categorize patients			**visit 1**	**visit 2**	**visit 2**	**visit 4**
Center 1	Treatment	Age/Sex/Race				
	Placebo	Age/Sex/Race	Responses subject to the effects o			
Center 2	Treatment	Age/Sex/Race	• Center,			
	Placebo	Age/Sex/Race	• Treatment,			
...	...	...	• Demographic classification			
			• Visit (the time factor)			
Center 10	Treatment	Age/Sex/Race	• Baseline measures			
	Placebo	Age/Sex/Race				

After baseline measurement, the patients in each center are randomly assigned to treatment and placebo and followed at four clinic visits. The patients' responses are affected by center, treatment, age, sex, race, the time of visit, baseline measures, and other factors unknown to this study. Of those known factors, center and treatment are the controlled factors. The patients are stratified by center, and they are randomly allocated to treatment groups. Visit is a chronological marker. Although the schedule is planned, visit is not a controlled factor because it has no bearing to patient assignment to treatment assignment. Visit potentially confounds with treatment effects if compliance to visit schedule is different between treatment groups. Baseline and demographics are covariates, i.e., they are

recorded uncontrolled factors. Because covariates are not controlled, patients divided by covariates are not necessarily comparable. For instance, there might be 10 male and 5 female patients in the treatment group in a center, and 5 male and 10 female patients in the placebo group. This uneven distribution of patients between gender groups renders gender a confounding factor in the evaluation of treatment effects in that center. If gender is added to the analysis of variance together with center and treatment, part of variations that would have been duly attributed to the effects of treatment will be wrongly attributed to the effects of gender. Analysis of variance with covariates is further addressed in Chapter Five, section 5.4.

4.2.2 Grouping and curve fitting

A categorical factor in the analysis of variance means that the patients are divided into groups by this factor, and the means and their standard errors are computed in these groups. A continuous factor in the analysis of variance means that the mean response curve is fitted over the continuous factor such that the sum of the squared differences between this mean curve and data is minimal. In statistical textbooks, this is often called least squares estimation. Simultaneous presence of categorical and continuous factors in the analysis of variance means that mean response curves are fitted in each category and then compared among the categories. Therefore, grouping and curve fitting are two basic operations in the analysis of variance.

Knowing the basic operations in the analysis of variance, researchers can plan analysis and make specifications to technical personnel. Much as using laboratory forms to order diagnostic tests, forms like what is shown in Table 4.3 may be helpful to order an analysis. Forms like that contain instructions on

- the identification of response variable, controlled factors, covariates and chronological marker,
- whether or not interaction effects or covariates are included in the analysis,
- whether the means are computed and then compared at each visit or the mean response curves over time are fitted and then compared as a whole, and finally
- how the results of analysis are presented.

The analysis specified in this table is to group the patients by center and treatment at each visit, compute the means and their standard errors, and present the result with an ANOVA table and a graphical display of the means. If the effects of center-treatment interaction is excluded, the analysis will be two separate groupings first by center and then by treatment. The effects of treatment are represented with the means of treatment groups, and the effects of center are represented with the means of centers.

Table 4.3 The Analysis of Variance Specification Form

Specifications for Analysis of Variance	
Response Variable:	blood sugar values
Controlled Factors:	center, treatment
• Interaction	center-treatment interaction
Covariates:	none
Chronological Marker:	
• Time-specific	yes, at each visit
• Curve over the time	
Presentation:	
• ANOVA table	yes
• Graphics	
• Means	yes
• Least squares means	treatment effects

The most conservative analysis is including only the controlled factors and attributing the unaccounted variations to the effects of the uncontrolled factors. The design of the study determines the logical validity of the analysis. Occasionally, it is profitable to adjust for the effects of covariates. Suppose we add sex, a categorical covariate, into the analysis. The result is that the patients in each center are first divided into groups of males and females, and treatment effects are then compared within each group of males or females. This adjustment for the effects of covariate is profitable if the distribution of gender is roughly balanced between treatment groups. If the distribution is unbalanced, adjusting for covariate would be a perfect way of introducing confounding, not present previously, into the evaluation of treatment effects. Adding continuous covariates results in mean response curves over the covariates, not just few means, being compared between treatment groups. Suppose we add

continuous baseline measure as covariate into the analysis. What the analysis does is to fit a straight line in each treatment group in each center and then compare the straight lines across treatment and center. The analysis of covariates is fully discussed in Chapter Five, section 5.4.

The effects of time on patients' response to treatment are best evaluated by comparing treatment effects at each visit. A line-scatter plot of the means over visit may be used to highlight the average response profiles. When visit schedule is irregular or there are abundant missing observations, however, timely comparison of treatment effects may not be feasible. In this situation, we may use mean response curves to represent the response profiles over the time and compare the whole profiles between treatment groups. In the above table, curve over the time is a call for this strategy of incorporating time into the evaluation of treatment effects.

4.2.3 Linear models as a technical language

Linear models are a technical language for the analysis of variance. Suppose the primary interest is the effects of baseline, center, treatment, and center-treatment interaction, and we would like to attribute all unaccounted variations to the effects of the uncontrolled factors. With linear model, this is simply

responses = baseline + center + treatment +
center-treatment interaction + residuals.

The actual variables corresponding to the specified effects are called explanatory variables. If an explanatory variable is continuous, its effects are represented by a mean response curve, known as regression curve; if it is categorical, its effects are represented by a group of means.

A linear model is essentially rules of mathematical manipulations for computing group means, mean response curves and their standard errors. The technical detail of general linear models is given in Appendix B. The use of mathematical distribution and the parameterization of linear models are entirely technical. Whatever rules set forth for a linear model are for the sole purpose of getting the desired summary measures, not the other way around. It is unfortunate that some rules of linear models are written

in statistical textbooks as "assumptions," which makes careless readers believe that those "assumptions" have to be validated prior to analysis. It is not uncommon that clauses on checking for model assumptions are written in research protocols; examples are checking for normal distribution, homogeneity of variance, additivity, and so on. Checking for these "assumptions" often requires making other assumptions, and the result is a logical loop that is going to nowhere.

4.3 Effects and their measurement

When multiple factors are evaluated simultaneously in the analysis of variance, the effects of any specific factor may be quantified with up to three different types of measures, depending upon how that factor is being evaluated in relation to others. The three types of measures are best explained with an example. Suppose two medical centers each recruit 155 patients, and the primary interest is the effects of center, treatment and center-treatment interaction. The data are summarized in the following table:

Table 4.4 Summary of Data in a Two-Center Trial

	Treatment A		Treatment B		Treatment C		Pooled	
Center	N	Mean	N	Mean	N	Mean	N	Mean
1	100	110	5	103	50	128	155	116
2	5	72	100	79	50	119	155	91
Pooled	105	109	105	80	100	124	310	104

A striking feature of the data is the uneven distribution of patients between treatments A and B across centers 1 and 2.

4.3.1 Type I measure

In essence, type I measure is simple averages without stratification. With type I measure, the effects of any factor are evaluated in isolation from the effects of other factors. The type I measure of the effects of treatment, for instance, is simply the means of pooled data across centers: A = 109, B = 80, and C = 124, as if the data were not stratified by center at all. A good thing about type I measure is that the data values are equally weighted, and they make fair contributions to the means regardless of

other factors. Problem is that the stratification by center in the study is not incorporated into the analysis, and consequently, the mechanism intended to control the confounding effects of center is not fully utilized. As shown in the summary table, most patients who received treatment A came from center 1, whereas most patients who received treatment B came from center 2. Therefore, the comparison of treatments A and B is also, more or less, a comparison of centers 1 and 2. In other words, quantified with type I measures, treatment and center confound the effects of each other.

The type I measure of the effect of center-treatment interaction is six essential contrasts among the means in the shaded area of the summary table: 110 – 103, 110 – 128, 110 – 72, 110 – 79, and 110 – 119, as an example. This measure of center-treatment interaction contains the effects of center and treatment in the sense that if the six means are similar, no effect can be claimed whatsoever. But if there is a difference, we have no clue from this measure whether the difference is due to the effects of center, treatment, center-treatment interaction, or any combination of the three. Because of this lack of specificity, type I measures are virtually useless for interaction effects.

Simple arithmetic may be all that is needed to compute type I measures. Of course, linear models can always be used for complex computations. With linear models, we put one factor or the interaction of multiple factors at a time. For instance, three linear models are needed to compute the type I measures for the effects of center, treatment and center-treatment interaction, with one factor or interaction at a time:

responses = center + residuals,
responses = treatment + residuals, and
responses = center-treatment interaction + residuals.

Some computational details are given in Appendix A. In general, type I measures are not desirable because they do not fully utilize the power of the analysis of variance technique.

4.3.2 Type II measure

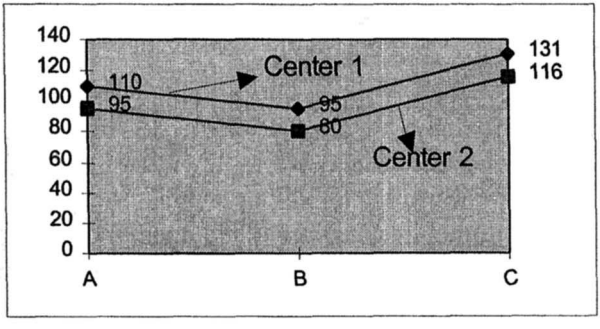
Type II measure is stratified averages representing the *additive* effects of involving factors. By additive effects, as opposed to joint or synergistic

effects which feature type III measures, we are making an assumption that the involving factors exert their effects on patients' responses independently. With type II measures, the effects of one factor are identical across other factors. As shown in the following table, for instance, the effects of treatment are the same in both centers:

Type II Measures

	A	B	C
Center 1	110	95	131
Center 2	95	80	116

From this table, A – B = 110 – 95 = 15 in center 1 is identical to A – B = 95 – 80 = 15 in center 2, and so are any other contrasts among the treatment groups. The graphical presentation is most characteristic for parallelism:



A linear model is generally required to compute type II measures. All we have to do is to put the factors of interest in the model without interaction effects. For instance, we may use

responses = center + treatment + residuals

to get the type II measures for the effects of center and treatment. More details on using linear models to compute type II measures are given in Appendix A.

Type II measure is averages weighted mostly by the number of observations across strata. The following table compares the means of treatment A in centers 1 and 2 to their type II counterparts:

Center	N	Mean	Type II
1	100	110	110
2	5	72	95

The two measures are consistent in center 1 where 100 patients were observed, whereas discrepancy shows in center 2 where only 5 patients were available. The assumption of additive relationship between center and treatment is blamed for this discrepancy. When the actual relationship is not perfectly additive, type II measures are adjusted to meet the restriction of parallelism. The actual computation is much like a lever; the groups with most of the patients dominate the balance, and groups with small number of patients are sacrificed to meet the requirement of additivity. While stratification distincts type II from type I measures, weighted averaging across strata by the number of observations distincts type II from type III measures.

4.3.3 Type III measure

Type III measure is stratified averages representing the *joint* effects of involving factors. Joint effects can be additive if the involving factors exert their effects independently. Joint effects are most interesting when they are significantly different from the additive effects of the involving factors. This is when the joint effects are better known as synergistic effects. The following table presents the type III measures for the joint effects of center, treatment and center-treatment interaction:

Type III measures

Center	A	B	C
1	110	103	128
2	72	79	119

They are actually the means in the shaded area of the summary table. The analysis may simply carried out with the linear model,

responses = center + treatment + center-treatment interaction + residuals.

Compared to their non-specific type I counterparts with 5 degrees of freedom, the type III measures for the effects of center-treatment interaction are cross-center contrasts of the within-center essential treatment contrasts:

A-B between C1 and C2: $(110 - 103) - (72 - 79)$, and
A-C between C1 and C2: $(110 - 128) - (72 - 119)$,

which have 2 degrees of freedom.

Type III measures preserve the rich information in the original data without any superimposed restrictions and present the joint effects of interrelated factors. Type III measures are good for evaluating heterogeneity. The effects of center-treatment interaction, for example, measure the differences of treatment effects across centers. Type III measures are good for searching for the combination of factors that generates the optimal therapeutic effects. The following table, for instance, clearly demonstrates the superiority of B + D combination over any other combinations:

	C	D
A	10	12
B	7	**31**

If the factors being studied have multiple levels, a grid surface of mean responses as shown in Figure 4.1 on the next page may be constructed in search for optimal responses. This three-dimensional plot of 121 mean responses represents the joint effects of drugs A and B at different concentration levels. This graph clearly shows that high mean responses generally associate with low concentrations of drug B. The effects of drug A depend on the level of drug B. At low levels of drug B, three peak responses are observed at high, medium and low levels of drug A, with the highest response at a low level of drug A. At high levels of drug B, the peak response occurs only at low levels of drug A. The only drawback of type III measures is that they may not be optimal for estimating the effects

of individual factors. The discussion is continued in section 4.4.2 on marginal means.

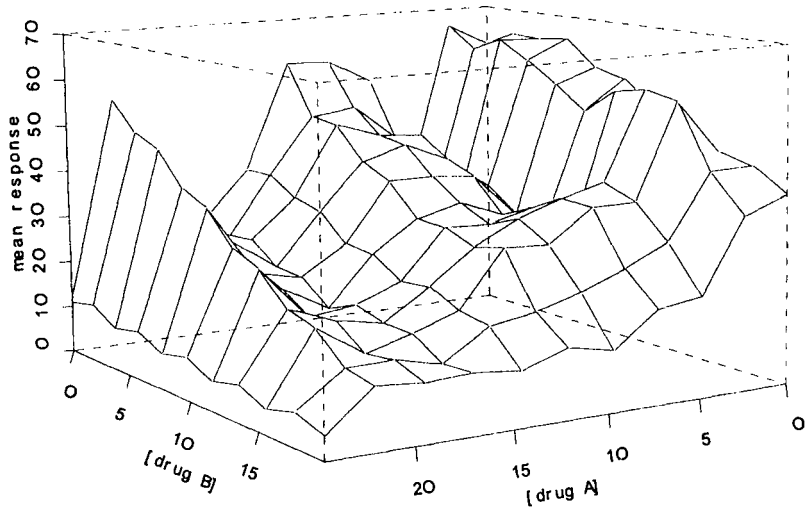


Figure 4.1 Mean response surface by concentrations of drugs A and B

4.3.4 Pros and cons and making choices

The pros and cons of the three types of measures are recapitulated in the following table:

Table 4.5 Features of Three Types of Measures

	Type I	Type II	Type III
Effects	Single	Additive	Joint
Weighting across strata	Yes	Yes	No
Stratification	No	Yes	Yes
Interaction	Limited use	No	Yes
Confounding with strata	Possible	Possible	Limited

Type I measures quantify the effects of a single factor in spite of its actual interaction with other factors. On one hand, type I measures represent the fair contributions of data values regardless of strata; on the

other hand, because more weight is given to strata with more patients, confounding from strata may be introduced when the number of patients is very different from stratum to stratum. Type I measures for interaction effects are not specific to the synergistic effects that most people refer to when speaking interaction.

Type II measures are also mainly used for evaluating the effects of individual factors. Compared to type I measures, the evaluation is stratified by other factors. The relationship among possible interrelated factors is assumed to be additive, and therefore, interaction effects are excluded by definition. The effects of a single factor, known as the main effects, are quantified with averages of data across strata weighted by the number of patients in each stratum. As type I measures, confounding from strata may be introduced when the number of patients is different from stratum to stratum.

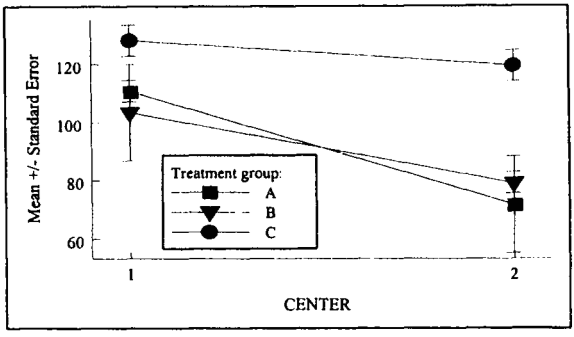
Type III measures are the measures of choice for studying the effects of interrelated factors unless it is known that those factors act additively on patients' responses. Because the data in each stratum stand alone, possible confounding from strata is curbed. Problem is when type III measures are combined to estimate the effects of a single factor, strata with quite different numbers of patients are equally weighted. This problem is further addressed in section 4.4.2 on marginal means.

4.4 Presentation of analysis results

The analysis of variance gives rise to results in different form and complexity, and the presentation requires careful consideration of purposes. Graphical display of the means and their standard errors is perhaps the best presentation. When the interest is the effects of individual factors, the least squares means are concise measures to quantify those effects. Finally, the mean sum of squares may be used to measure the average effects of a class of factors. From the means to least squares means and to mean sum of squares, there is a continuous loss of information in exchange for conciseness. Although statistical testing is widely advocated in the statistical academia and adopted in the current research practice, it has absolutely *no* role in the analysis of variance here whatsoever.

4.4.1 Graphical presentation

Graphical display of means and their standard errors is perhaps the most informative way to present the results of analysis of variance. The following line-scatter plot presents the type III means for the joint effects of center and treatment:



The means are computed with the linear model,

responses = center + treatment + center-treatment interaction + residuals.

It appears that the mean responses to treatment C are significantly higher than those to treatments A and B in both centers. The mean responses to treatments A and B are similar. The differences among treatments in center 1 are somewhat smaller than the differences in center 2.

Graphical presentation is a necessity when analysis involves continuous explanatory variables. The result of analysis is mean response curves, and the number of means directly depends on the number of unique values of the continuous explanatory variables. Presenting mean response curves as opposed to voluminous numeric means makes it easy to grasp the information. In the analysis specified in the linear model,

responses = baseline + center + treatment +
center-treatment interaction + residuals.

baseline is a continuous variable. The analysis results in about 150 means representing the joint effects of baseline, center, and treatment. If numerically listed, these 150 means would be a great challenge to human eyes and mind. It becomes a breeze to sort out the underlying information when they are nicely presented together with their standard errors on the following graph:

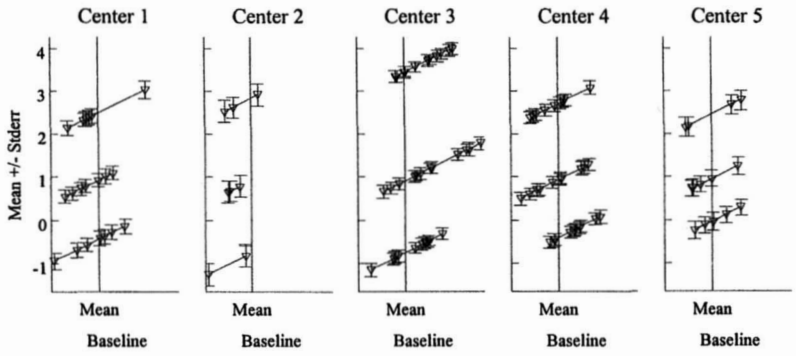


Figure 4.2 Mean response curves over baseline by center and treatment

In this graph, each line represents the mean responses over baseline values in each treatment group in each center. There are significant treatment effects because the lines are well separated, and the treatment effects are consistent across all centers. For computing least squares means, the points of interest are where the vertical line at the mean baseline crosses each oblique line in each center. These points represent the mean responses in different treatment groups and centers at the mean baseline.

4.4.2 Marginal means or least squares means

When multiple factors are evaluated simultaneously with the analysis of variance, the mean responses may be combined to estimate the effects of individual factors. A way of combination is averaging the mean responses, and the resulting means are referred to as marginal or least squares means (LSM).

It is straightforward to compute least squares means if all explanatory variables are categorical. For the analysis specified in this linear model,

responses = center + treatment + residuals,

there are six type II means representing the additive effects of center and treatment, and they are tabulated in the shaded area:

Treatment:	**A**	**B**	**C**
Center 1	110	95	131
Center 2	95	80	116
LSM	**102**	**87**	**124**

The least squares means for the effects of treatment alone are simply the average of those type II means in each treatment group over centers:

$$\text{A: } \frac{110+95}{2}=102, \quad \text{B: } \frac{95+80}{2}=87, \quad \text{and C: } \frac{131+116}{2}=124\,.$$

The least squares means derived from type II means can be quite different from those derived from type III means. By the additive nature of type II means, the effects of treatment are identical across centers. Therefore, the contrast of least squares means between any two treatment groups is identical to the corresponding contrast in every center: A – B = 102 – 87 (LSM) = 95 – 80 (Center 2) = 110 – 95 (Center 1) = 15. This is not necessarily true for the least squares means derived from type III means. From the analysis specified in the linear model:

responses = center + treatment + center-treatment interaction + residuals.

the following table presents the type III means for the joint effects of center and treatment:

Center	A	B	C	LSM
1	110	103	128	114
2	72	79	119	90
LSM	**91**	**91**	**124**	

The least squares means of center are

$$\frac{110+103+128}{3} = 114 \text{ and } \frac{72+79+119}{3} = 90,$$

and the least squares means of treatment are

$$\frac{110+72}{2} = 91, \quad \frac{103+79}{2} = 91 \text{ and } \frac{128+119}{2} = 124.$$

The contrast of least squares means between treatments A and B is 91 – 91 = 0 while its within-center counterparts are 110 – 103 = 7 in center 1 and 72 – 79 = –7 in center 2. This inconsistency is due to the effects of center-treatment interaction. The effects of treatment are different from center to center, and the least squares means of treatment represent a combination of heterogeneous effects over centers. In the presence of significant center-treatment interaction effects, the least squares means of any single factor can be very misleading because they are averages over a hodgepodge of measures that do not belong to the same category.

The least squares means from type II measures are generally weighted with the numbers of observations, while the least squares means from type III measures are not. By weighting with the number of observations, the least squares means are a fair representation of contributions from individual data values. The shortcoming is possible introduction of confounding from unbalanced patient distribution over strata. For example, of the 105 patients who contribute to the least squares mean of treatment A, 100 are from center 1 while only 5 from center 2. Thus, the least squares mean more represents the effect of center 1 than that of center 2. The least squares mean of treatment B is the opposite. Consequently, the effects of center more or less confound the effects of treatment. Confounding of this kind is confined with type III measures. The type III means are specific to both treatment and center, and because there is no averaging across strata, they are independent of each other. This independence confines in some degree the effects of confounding, if any, between treatment and center. The least squares means are averages with an equal weight. The problem with this equal weighting is that the quality of type III means is not taken into account. For example, the type

III mean from 5 patients generally has poor reliability and robustness than that from 100 patients. Indiscriminate combination of type III means of different quality may jeopardize the quality of resulting least squares means.

If the analysis involves continuous explanatory variables, the current practice is to use the mean response at the means of continuous explanatory variables to represent their effects. Although it leads to the smallest variance, choosing the mean is entirely a convention. Once the mean responses at the means of continuous variables have been determined, the continuous variables may be viewed as categorical variables, and one may proceed, as usual, to compute the marginal or least squares means.

4.4.3 The analysis of variance table

A less informative summary than graphics is the analysis of variance table, known as ANOVA table, in terms of mean sum of squares. Sum of squares, mean sum of squares and degree of freedom are defined in section 4.1. The following table summarizes the result of the analysis specified in this model,

responses = center + treatment + center-treatment interaction + residuals:

ANOVA Table

Source	*DF*	*SS*	*MS*	*F*	*P*
Center	1	11326	11326	8.22	0.0044
Treatment	2	29157	14579	10.59	0.0001
Interaction	2	4147	2073	1.51	0.2235
Residual	304	418625	1377		

SS: sum of squares, MS: mean sum of squares,
DF: degree of freedom, F: ratio over residual MS, P: p-value

Two unfamiliar elements in this table are the F and P values. F value or F statistic is the ratio of mean sum of squares for the effects of interest over the residual mean sum of squares. A large F value indicates that the variation caused by the effects is greater than that caused by the effects of the uncontrolled factors, and therefore, suggests a significant effect. P value is a map of F statistic to the scale of 0 to 1, [0, 1], by comparing to

an F distribution. As a measure, p-value is equivalence to the corresponding F value. The only difference is scale. A general discussion on mathematical distribution and P value is given in Chapter Ten. At this point, one must not fantasize F and p values, but view them as alternatives, preferred by some, to the ratio of mean sums of squares.

With ANOVA table, we sacrifice detailed information for conciseness. The mean sum of squares is not specific to any particular comparison of means; rather, it is an agglomerate measure representing the average of a class of pairwise comparisons of means. A large contrast may be dampened when averaged together with a large number of small contrasts and could be overlooked if the resulting mean sum of squares is not significant. On the other hand, a large mean sum of squares by no means guarantees that all the constituent contrasts are equally large because the contributions from few large contrasts may just be sufficient to make the mean sum of squares significant.

4.5 Limitations

The analysis of variance technique has limitations. It is well known that the mean is subject to overdue influences from few outlying data values and is completely useless when the data distribute in clusters. However, more critical is the use of standard error, which generally decreases in proportion to the number of observations and may underestimate the effects of the uncontrolled factors. Further discussion on the mean and standard error are given in Chapter Three.

The standard error of in analysis of variance has its own problem that it is not specific to the mean itself. The residual mean sum of squares is the common numerator for the standard errors of all the means. That residual mean sum of squares measures the average deviation of all the means, not any specific mean, from the original observations. Therefore, the standard error of a mean in the analysis of variance may not be a good measure of the quality of that mean.

4.6 Pairwise comparisons of multiple means

In recognition that the standard error in the analysis of variance is not specific to the mean, pairwise comparisons of individual means involving only the observations in the comparing groups may be preferred,

especially when the quality of comparisons is quite different from pair to pair. Simultaneous pairwise comparisons of means are best made visually with graphs. For even a small number of means, the number of pairwise comparisons can be formidable. By graphical presentation, however, a large number of means can be condensed on a single page, and conspicuous features can be quickly spotted. Suppose two medical centers recruit 90 patients and the data are summarized in the following table.

Summary Table

	Treatment A			Treatment B			Treatment C		
Center	N	Mean	Std	N	Mean	Std	N	Mean	Std
1	16	2.759	0.75	17	2.568	1.00	17	2.912	0.76
2	13	2.459	0.70	14	2.646	0.92	13	2.665	0.71
Pooled	29	2.625	0.73	31	2.604	0.95	30	2.805	0.74

To evaluate the effects of center, treatment, and their interaction, 23 pairwise comparisons need to be made. These laborious comparisons are largely saved by the following graph:

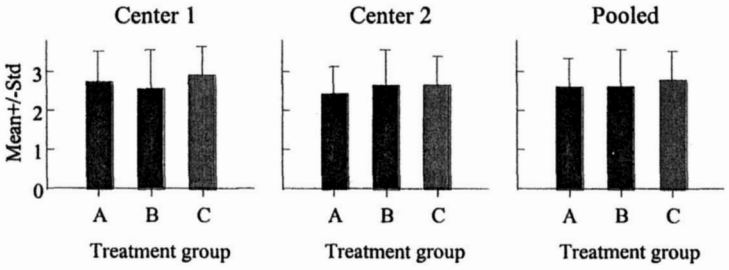


Figure 4.3 Pairwise comparisons of means and their standard deviations

which delivers a clear message that the mean responses are similar in both centers.

The analysis of variance and simultaneous pairwise comparisons of means are two different approaches of data analysis, and it is not surprising, therefore, that the results may disagree. Pairwise comparisons

are specific and sometimes preferred to the analysis of variance. Of course, when multiple factors are evaluated simultaneously, the number of pairwise comparisons can be prohibitively large. In that situation, the analysis of variance, offering measures that summarize the data in various detail, may be used for an overview of the data.

It is scientifically profitable to make all possible comparisons of observations. The concept widely spread in the statistical academia that a conclusion drawn from multiple comparisons is more prone to error than one from a single comparison is baseless. This so-called issue of multiplicity is analyzed and criticized in Chapter Ten. The only practice that should be avoided is selecting the results of favorite comparisons and generalizing them. In general, comparisons based on subset of data may not have sufficient quantity of observations to warrant an adequate control of confounding. The results of the selected comparisons may be simply caused by an uneven distribution of the uncontrolled factors between the comparing groups. By averaging over a large number of observations, the impact of any single uncontrolled factor is proportionally dampened. Therefore, when the analysis of variance based on whole set of data shows no significance even though some individual comparisons do, researchers should exercise caution before arriving at any hasty conclusion.

4.7 The analysis of variance for categorical data

Table 4.6 Summary Table of Categorical Data in a Trial

		Center 1		Center 2		Center 11	
Category	Treatment	Mean	N	Mean	N	Mean	N
Up	Drug	0.0139	72	0.0250	80	0.2018	109
	Placebo	0.1500	40	0.0000	40	0.0833	60
Down	Drug	0.9028	72	0.8875	80	0.7156	109
	Placebo	0.7500	40	0.1000	40	0.7500	60
Zero	Drug	0.0833	72	0.0875	80	0.0826	109
	Placebo	0.1000	40	***0.0000***	40	0.1667	60

If categorical responses are coded with 1 and 0, the mean for a specific category is the percentage of data coded with 1. As long as the mean is used to summarize the data, the analysis of variance for categorical data can be proceeded as if the data were continuous. This practice generally

produces acceptable results for categorical data. The only difference is that for polychotomous responses, the same analysis needs to be repeated for all the categories.

Suppose in a multicenter trial the response of any particular patient falls in one of three mutually exclusive categories: up for increase, zero for no change and down for decrease. The data are summarized in Table 4.6 on the previous page for three selected centers. Because the responses are trichotomous, the effects of center, treatment, and their interaction need to be evaluated in each category. Therefore, three similar analyses need to be run, and they may be specified with the linear models,

up, down or zero = center + treatment +
center-treatment interaction + residuals.

The type III means from the analysis are identical to those means in the summary table except for center 2 where the mean of ZERO for placebo is -0.0000 from the analysis of variance.

Regular linear models are simple, practical and generally acceptable for categorical data; but they are not technically perfect. The means from the analysis of variance may go out of the range slightly especially when the observed frequencies are close to 0 or 1. An example is that negative mean of –0.0000, while the observed frequency is 0. This problem can be easily solved by performing the analysis on a scale that does not admit negative values; for instance, the logarithmic scale:

log (up, down or zero) = center + treatment +
center-treatment interaction,

The logit scale is well accepted for the analysis of variance on categorical responses. With the logit scale, the mean responses are guaranteed to be within [0, 1]. Linear models on the logit scale are called logistic models. Chapter Six will discuss the analysis of variance on an arbitrary scale, where logistic models, log-linear models and etc. are special cases. The analysis of variance on categorical data can largely replace the widely taught chi-square tests for contingency tables and the commonly used Mantel-Haenszel test in epidemiological textbooks.

This Page Intentionally Left Blank

5

Frequently Debated Issues in the Analysis of Clinical Trial Data

Summary

The issues discussed in this chapter are frequently encountered in the analysis of clinical trial data. There are no universally agreed solutions to these issues, and they are often open to debate in scientific forums or public hearings. The first issue concerns the effects of center-treatment interaction. When the number of observations in each center is small, the center-treatment interaction effects may be evaluated by comparing the trends of treatment effects across centers. However, if the interaction effects and the residual sum of squares cannot be reliably estimated at the same time, the interaction effects should not be claimed. The second issue concerns adjustment for the effects of covariates. Comparisons of treatment effects may be improved by appropriately attributing some variations to the effects of covariates. However, covariates must not confound the effects of treatment for a profitable adjustment. The third issue concerns end-point analysis versus analysis over the time course. It is most informative to compare treatment effects over time, and it is technically advantageous when there are abundant missing data.

5.1 An overview

This is a typical clinical trial that multiple centers are initiated simultaneously, patients are screened and baseline measures are taken during the run-in period, eligible patients in each center are randomly assigned to treatment groups, and the randomized patients are followed in a series of planned visits. An example was given in Chapter Four, section 4.2.1, where a table was designed to visualize the structure of such a trial.

Medical practice can be quite different from center to center, and indeed, the effects of center generally cause a great deal of data variation. Because of the high cost of clinical studies, however, instead of reducing the number of centers, the current designs continue the trend of initiating many study centers simultaneously and allocating to each center a very limited patient recruitment quota. This practice speeds up patient enrollment and reduces the impact of individual centers on the business process and final result. The drawback is the lack of sufficient observations to evaluate the consistency of treatment effects across centers. In the analysis of variance, center-treatment interaction measures the consistency of treatment across centers. The first two sections will focus on the evaluation of center-treatment interaction.

Chapter Four, section 4.2.1 explains the importance to distinguish between the controlled and uncontrolled factors. Controlled factors are those that determine the assignment of patients to treatment groups, while uncontrolled factors have nothing to do with patient assignment. Analysis with only the controlled factors fully utilizes the mechanism of study design, such as randomization and stratification, for the control of confounding. Nevertheless, some uncontrolled factors, such as baseline measures and patient demographics, often have significant effects on the responses. Stratification by these factors in analysis creates homogeneous patient groups in which comparisons of treatment effects will have a better precision. Recorded uncontrolled factors are collectively called covariates. Section 5.4 will discuss the analysis of variance with covariates.

Currently, the reporting of clinical studies is still by and large based on endpoint analysis, the analysis of data collected only at the end of study. There is no doubt that endpoint analysis greatly simplifies the

reporting process, and often, the endpoint result is what many people care about after all. It is important to realize, however, that endpoint analysis has serious problems. Conceptually, no one knows for sure when is the end. When knowledge *a priori* is extremely limited, endpoint is usually defined arbitrarily. One point may be just as good as another. Technically, a clear-cut endpoint is difficult to define for a study because a 15% withdrawal rate is generally expected in even a well tolerated and well controlled study. If we define end point as the planned end of the study, those patients lost in follow-up will not contribute any information to the analysis. To recover the information from patients lost in follow-up, a popular definition of end point is the last observed responses at or before the planned end of the study, the so-called last-observation-carried-forward (LOCF) approach. This approach neglects the time effect on responses. The resulting data for analysis are a mixture of both early and late responses. Section 5.5 proposes an alternative to end-point analysis. The philosophy is presenting response profiles over time.

Mean responses have been the primary interest in the analysis of clinical trial data. Individual responses not fully represented by the means are, however, equally important and deserve careful evaluation. Examination of residuals is effective for identification of far-from-average individuals. Section 5.6 is devoted to residual analysis that has long been overlooked.

5.2 No center-treatment interaction

Evaluation of center-treatment interaction generally involves comparisons of treatment groups between centers. As a simple example, for the data summarized in the following table,

	Center 1	Center 2
Treatment A	M_{A1}, N_{A1}	M_{A2}, N_{A2}
Treatment B	M_{B1}, N_{B1}	M_{B2}, N_{B2}

where M and N denote the mean and number of observations in each treatment group in each center, the effects of center-treatment interaction are measured by the contrast,

$$(M_{A1} - M_{B1}) - (M_{A2} - M_{B2}),$$

which is the difference of within-center treatment contrasts between centers.

In multicenter trials where each center recruits only few patients, the claimed effects of center-treatment interaction are often based on the means of insufficient observations. In the previous example, the N's can be as small as 1 or 2, and the M's are the means of those 1 or 2 data values. Those means are almost as volatile as a single observation against the uncontrolled factors. Hence, the claimed interaction effects by and large confound with the effects of the uncontrolled factors. In this circumstance, the effects of center-treatment interaction and the effects of the uncontrolled factors cannot be truly separated and reliably estimated at the same time, and a conservative approach is not to claim the effects of interaction, but ascribe them to the effects of the uncontrolled factors.

Suppose 20 centers are initiated for a trial, and each is budgeted for 8 patients being assigned randomly to 4 treatment groups. Although it is technically feasible to claim the effects of center-treatment interaction by the analysis,

responses = center + treatment + center-treatment interaction + residuals,

which is summarized in the following table:

ANOVA Table

Source	*DF*	*SS*	*MS*	*F*	*P*
Center	18	23.89	1.33	1.54	0.10
Treatment	3	0.51	0.17	0.20	0.90
Interaction	54	41.91	0.89	1.03	0.46
Residual	69	59.63	0.86		

SS: sum of squares, MS: mean sum of squares,
DF: degree of freedom, F: ratio over residual MS, P: p-value

The claimed effects of center-treatment interaction are based on the means of at most two observations in each treatment group in each center, and the mean sum of squares measures the average of 54 essential contrasts of those means. It would be difficult to convince people that the means of

two observations are robust against the effects of the uncontrolled factors. The interaction effects might just be a misclaim of what would have been the effects of the uncontrolled factors. It would be more reasonable to give up the claim and ascribe the variation to its true source. This may be done by fitting the main effects model,

$$\text{responses} = \text{center} + \text{treatment} + \text{residuals},$$

which is summarized in the following table:

ANOVA Table

Source	*DF*	*SS*	*MS*	*F*	*P*
Center	18	25.43	1.42	1.62	0.07
Treatment	3	0.64	0.21	0.24	0.87
Residual	123	107.53	0.87		

SS: sum of squares, MS: mean sum of squares,
DF: degree of freedom, F: ratio over residual MS, P: p-value

The sum of squares that was claimed as the effects of center-treatment interaction is now part of the residual variation, and the residual mean sum of squares is now more reliably estimated with 123 essential contrasts.

There are no hard-and-fast rules to determine when the effects of interaction should not be claimed. Strong center-treatment interaction generally requires fewer observations to show its effects than interaction of borderline significance. It is often an occurrence that the number of observations in each treatment group in each center is moderate, and analyses with and without center-treatment interaction show noticeable differences. Then it is a matter of judgment whether or not the effects of center-treatment interaction really exist and can be reliably estimated.

5.3 The effects of center-treatment interaction

In multicenter trials, significant center-treatment interaction suggests that the effects of treatment are different from center to center. Whenever possible, the effects of center-treatment interaction should be evaluated to demonstrate the consistency of treatment effects across centers. However, to speed up patient enrollment and minimize the impact of individual centers on the overall results, the current practice tends to initiate a large

number of investigation centers simultaneously and allows each center to enroll only a small number of patients. When the number of patients in each center is small, the means in each treatment group in each center can be as volatile as individual data values against the uncontrolled factors, and the variation of these volatile means cannot be claimed to be any but the effects of the uncontrolled factors. So the question is how to evaluate center-treatment interaction when the number of observations in each center is not large.

When direct comparisons of group means are not reliable, an option is to compare the trends of treatment effects across centers. Although comparisons of trends may not be as informative and sensitive as direct comparisons of group means for detecting the effects of center-treatment interaction, trends may be more reliably estimated by averaging all available observations in each center, whereas the already insufficient number of observations in each center have to be broken down by treatment in order to get group means.

Suppose 20 centers each recruit three patients who are then assigned to drugs A, B, and placebo. There is only one patient in each treatment group in each center. The analysis for comparing treatment effects and the linear trends of treatment effects across centers can be specified as

responses = center + treatment + center-treatment interaction + residuals,

where treatment is coded with 1 for drug A, 2 for drug B, and 3 for placebo and viewed as a continuous variable. The coding of treatment groups is completely technical, and the order has no meaning. When treatment is viewed as a continuous explanatory variable, the center-treatment interaction in this model defines a group of straight lines, and the slopes of these lines are compared across centers. The result of this analysis is pictorially presented in Figure 5.1 in the following page, where each line represents the trend of treatment effects in a center. The treatment effects in four centers shown in the upper plot are significantly different from the majority of centers. Down trend is shown in roughly 50% of the centers, and up trend in the other 50%. The overall picture does not seem to support the claim of consistent treatment effects across centers.

When treatment is coded into a continuous explanatory variable, the shape of trend must be defined in the analysis. By default, a continuous explanatory variable in a linear model defines a straight line. A defined shape may be too rigid to fit the data. For instance, when the response profiles are better described with a curve, fitting a straight line may not capture the signal. In fact, speculating a shape for the response curve is required in the analysis of variance with any continuous explanatory variables, and there is always a risk for lack of fit. The pain is partially relieved with polynomial curves, which allow for more flexibility when straight lines are not adequate. Graphical exploration of the data is also helpful to specifying an appropriate shape for the response curve.

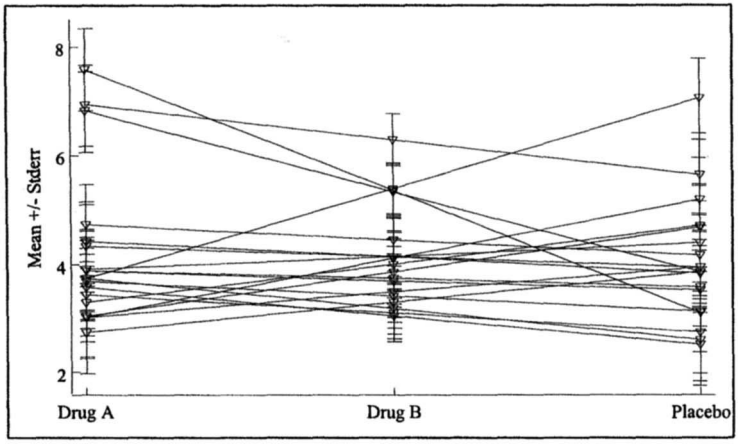


Figure 5.1 Linear trend of mean responses across centers

5.4 The analysis of variance with covariates

Covariates are factors or explanatory variables that are observed and recorded, but not used to guide the assignment of patients to treatment groups. Chapter Four, section 4.2 introduces the use of covariates in the analysis of variance. In adjustment for the effects of covariates, the basic operation is stratification if the covariates are categorical or fitting of curves if they are continuous. The purpose is to improve the comparisons of treatment effects by attributing some of the variations to the effects of covariates.

5.4.1 Baseline measures

Baseline is the measurement of response parameters before the administration of treatment. Baseline measures are separated from other covariates that are not direct measurement of response parameters. Together with patients' demographic measures, baseline may be compared among treatment groups to evaluate the effectiveness of randomization. In crossover trials (see Chapter Eight), baseline measures are extremely important in the evaluation of crossover effects. Another important use of baseline is to improve the precision of the comparisons of treatment effects by explaining some variations of responses, and this is the primary interest in this section.

The effects of baseline may be adjusted for in three different ways. The first is change from baseline:

change from baseline = response measure after treatment - baseline.

Change from baseline is ideal when responses to treatment are independent of baseline. For example, if all patients on treatment gain 30 pounds regardless of their baseline body weights, change from baseline for these patients would be all 30, and thus, the variation due to baseline is totally eliminated. However, if overweight patients tend to gain more weight than patients of normal weight, percent change from baseline is more appropriate:

$$\text{percent change from baseline} = \frac{\text{measure after treatment - baseline}}{\text{baseline (if} \neq 0)}.$$

The ideal situation for percent change from baseline is when responses to treatment are proportional to baseline.

A note of caution is that both change from baseline and percent change from baseline may fail to improve comparisons and even introduce additional variations if the underlying relationship between response and baseline does not warrant either of the adjustments. A good practice is to explore the relationship with, for instance, graphical techniques before attempting any adjustment.

The third is statistical adjustment and will be discussed next in the context of analysis of variance with covariates.

5.4.2 Statistical adjustment for the effects of covariates

In clinical studies, baseline measures, medical history, prior treatment as well as demographic information of the patients are the commonly encountered covariates. Appropriate adjustment for their effects may significantly improve the comparisons of treatment effects.

Adding categorical covariates into the analysis stratifies the patients so that treatment effects are compared within the strata. Suppose three treatment groups are compared in a six-center trial. The analysis without adjustment for the effects of gender is specified in the model,

$$\text{responses} = \text{center} + \text{treatment} + \text{residuals},$$

and the results are summarized in the following table:

ANOVA Table

Source	*DF*	*SS*	*MS*	*F*	*P*
Center	5	7.9472	1.59	1.07	0.38
Treatment	2	0.5731	0.29	0.19	0.82
Residual	**134**	**199.14**	**1.49**		

SS: sum of squares, MS: mean sum of squares,
DF: degree of freedom, F: ratio over residual MS, P: p-value

The least squares means of treatments are presented in the following table:

	LSM	STDERR
Treatment A	2.46	**0.20**
Treatment B	2.30	**0.18**
Treatment C	2.31	**0.36**

Now we adjust for the effects of gender by adding sex into the model:

$$\text{responses} = \text{center} + \text{sex} + \text{treatment} + \text{residuals}.$$

By adding sex, the patients in each center are stratified by gender, and then the effects of treatment are compared within each gender group. The improvement in precision is reflected on the reduction of residual mean sum of squares and the standard errors of the least squares means:

ANOVA Table

Source	*DF*	*SS*	*MS*	*F*	*P*
Center	5	10.95	2.19	2.49	0.0342
Gender	1	82.27	82.27	93.62	0.0001
Treatment	2	0.011	0.006	0.011	0.9940
Residual	**133**	**116.9**	**0.879**		

SS: sum of squares, MS: mean sum of squares,
DF: degree of freedom, F: ratio over residual MS, P: p-value

and

	LSM	STDERR
Treatment A	2.40	**0.16**
Treatment B	2.39	**0.14**
Treatment C	2.36	**0.28**

It is not always profitable to adjust for the effects of categorical covariates. As a principle, covariates must not confound with the effects of treatment. Suppose treatment group T has 10 male and 2 female patients while treatment group P has 2 male and 10 female patients:

	T	P
Male	10	2
Female	2	10

More male patients contribute to the effects of T, and more female patients contribute to the effects of P. Therefore, the effects of treatment more or less confound with the effects of gender. If we adjust for the effects of gender when comparing the effects of treatment, we might just mistakenly attribute to the effects of gender some variations that should have been attributed to the effects of treatment. Another situation where adjustment for the effects of categorical covariates is not beneficial is when the covariates cause too much reduction in the degree of freedom but not

enough reduction in the residual sum of squares. This over-consumption of degree of freedom may actually inflate the mean residual sum of squares. Although this has to do with the analysis of variance technicality, the underlying problem may well be that the effects of covariates are too trivial comparing to the effects of other uncontrolled factors. In circumstances like this, one may stop claiming for the effects of covariates and return the covariates back to the pool of uncontrolled factors.

Adding continuous covariates into the analysis defines mean response curves, and the effects of treatment are evaluated by comparing response curves instead of static means. The following analysis adjusts for the effects of baseline:

responses = baseline + center + treatment + center-treatment interaction + residuals.

The means and their standard errors line up in lines and are presented in the following graph:

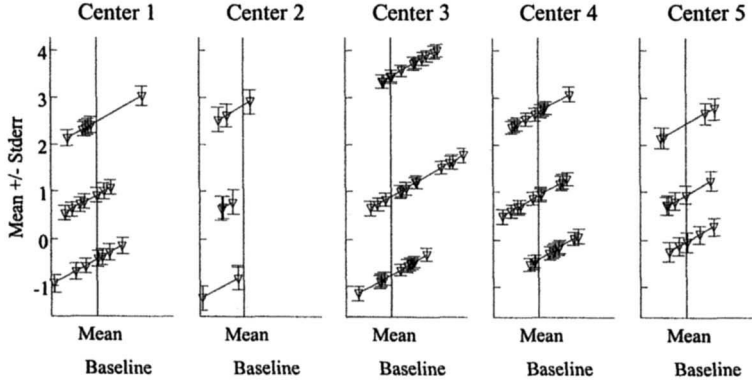


Figure 5.2 Mean responses over baseline measures by center and treatment

The slope of these lines represents the effects of baseline, and the variation along these lines is accounted for with baseline, which would otherwise

have been part of the residual sum of squares ascribed to the effects of the uncontrolled factors.

Statistical adjustment for the effects of baseline can be misleading if the baseline levels are uneven between treatment groups, or the treatment tends to have different effects at different baseline levels. The latter is known as baseline-treatment interaction. If the baseline level in treatment group is higher than the level in placebo group, for instance, the analysis adjusted for the effects of baseline may falsely attribute the variations truly caused by treatment to the effects of baseline. Therefore, a check for equal distribution of baseline between comparing groups is necessary. Nonetheless, for large, well-designed trials, it is generally safe to adjust for the effects of baseline measures, knowing that effective randomization guarantees equal distribution of baseline measures.

As with all continuous explanatory variables in the analysis of variance, the shape of mean response curves must be defined in the analysis. Whatever a shape defined in analysis is a speculation of the data, and chances are that the shape may not fit the data and result in loss of information. By default, continuous covariates in linear models define straight lines. When straight lines appear to be too restrictive, polynomial curves are much more flexible. The next section will show how to specify polynomial curves in the analysis of variance.

5.5 Mean response profiles in a time course

When patients are followed up for a period of time, their response profiles over the time course are generally much more informative than their responses at any particular moment. This section discusses two basic techniques to characterize mean response profiles to treatment over time.

Comparison of mean response profiles with the analysis of variance depends upon the schedule of data acquisition. If the schedule is regular with a narrow window, such as every 7 ± 3 days, the responses may be analyzed separately at every time point, and the results are simultaneously presented over the time points. An example is comparing the mean response profiles of an efficacy parameter from a multicenter trial. Because the visit schedule is regular, timely comparisons of treatment effects are feasible. The mean response profiles are obtained by first repeating the same analysis specified in this linear model,

scores = center + treatment + center-treatment interaction + residuals,

at every visit and then plotting the least squares means and their standard errors for treatment against the scheduled time:

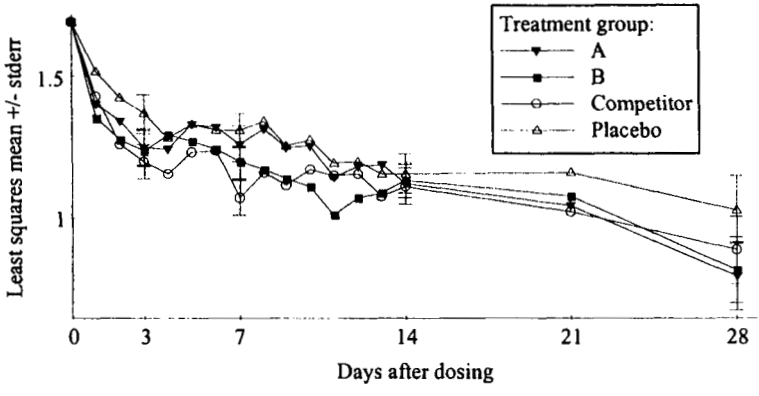
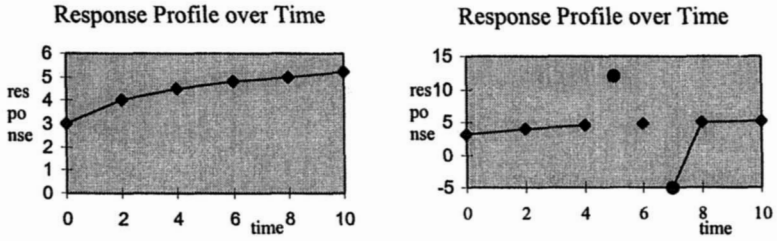


Figure 5.3 Mean response profiles over time by treatment

Although at some points of time, there are some differences among the four treatment groups, overall, the study failed to demonstrate any significant therapeutic advantage over placebo. This study demonstrated a significant placebo effect. A good deal of improvement is observed in the placebo group over time.

If the schedule is irregular or missing visits are abundant, it would be impossible to make comparisons within a common time frame. A solution is to fit mean response curves in each treatment group and then compare the curves among treatment groups. Comparing patterns instead of points is always an effective strategy for handling missing data. This is because a pattern may be perceived from only few observations. For instance, if the response profile is adequately represented with a straight line over time, in principle, two observations are sufficient to estimate the line, no matter how many other observations are missing. This is consistent with our geometrical experience that two points make a line, and three points could make a curve. However, the assumption is that there are no capricious

changes from point to point, and the transition from point to point is smooth. This is illustrated in the following graph:



The smooth link through the responses at time 4, 6 and 8, shown in the left graph, is based on the assumption that those three responses can roughly predict the responses between 4 and 8. If the responses obtained later at 5 and 7 turn out to be black sheep, as shown in the right graph, then the beautiful theory is brutally destroyed by the ugly fact. Nevertheless, "God is subtle, not capricious." If the time interval is reasonably small, smooth link through discrete points is seldom upset by surprises.

Suppose in a single center trial the visit schedule is irregular and missing values are abundant. The linear model on the identity and logarithmic scales specifies the analysis for comparing the frequencies of disease progression between treatment groups over time:

$$\begin{matrix}\text{mean(freq)} \\ \log[\text{mean(freq)}]\end{matrix} = \text{poly(time,2)} + \text{treatment} + \text{poly(time,2)} - \text{treatment interaction},$$

where poly(time, 2) denotes a quadratic polynomial curve over the time points:

$$\text{poly (time, 2)} = a \cdot \text{time} + b \cdot \text{time}^2,$$

a and b are the coefficients to be determined. In this model, poly(time, 2) represents the combined mean polynomial curves over treatment groups; treatment represents the combined mean responses in each treatment group over time points; most interesting is the effects represented by poly(time, 2)-treatment interaction, which is the difference of mean

response profiles between treatment groups. The form introduced in Chapter Four, section 4.2.2 is actually better to specify this analysis clearly:

Specification for Analysis of Variance	
Response Variable:	frequencies of progression
• Scale	identity and logarithm
Controlled Factors:	treatment
• Interaction	yes
Covariates:	
Chronological Marker:	
• Time-specific	
• Curve over the time	quadratic polynomial
Presentation:	
• ANOVA table	yes
• Graphics	
• Means	yes
• Least squares means	

The identity scale means that we compare the mean responses directly; the logarithmic scale means that we compare the logarithm of mean responses. The logarithmic scale is used to avoid negative means for frequency data. The means and their standard errors are summarized in the following charts:

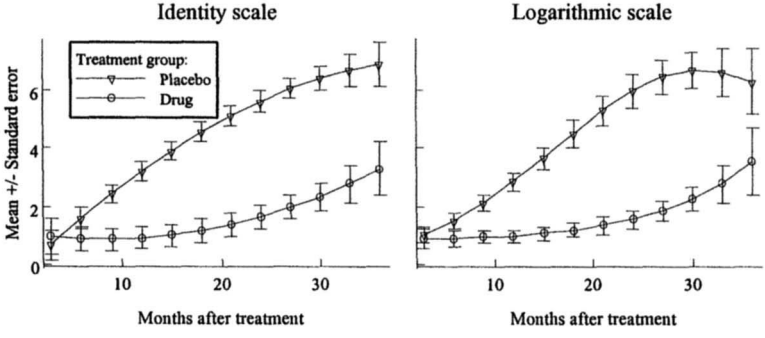


Figure 5.4 Mean response profiles over time by treatment

The result of the analysis is also summarized in this ANOVA table:

ANOVA Table

Source	*DF*	*SS*	*MS*	*F*	*P*
Time	2	68.41	34.21	1.75	0.17
Treatment	1	27.94	27.94	1.52	0.22
Interaction	2	188.87	94.44	5.12	0.006
Residual	**564**	10395	18.43		

SS: sum of squares, MS: mean sum of squares,
DF: degree of freedom, F: ratio over residual MS, P: p-value

In this analysis, the most desirable comparisons are represented by the time-treatment interaction, which measures the difference of mean response profiles between treatment groups. Collapse over time makes treatment effects not time-specific; collapse over treatment makes time effects virtually useless for comparing the effects of treatment. The analysis clearly demonstrates a significant therapeutic advantage over placebo growing over the time course.

Notice that the degree of freedom for the residual sum of squares is 564 while there are only 80 patients in the study. What happened is that the number of observations, not the number of patients, was used to compute the degree of freedom. A criticism to this is that observations from a single patient are not distinguished from observations from different patients, and when a large number of observations are actually obtained from very few patients, the observations do not represent the patient population. Some statistical authorities insist that a "repeated measures" type of analysis be more appropriate, in which the responses from each patient are viewed as a unit and the number of patients, not observations, is used to measure the strength of evidence. "Repeated measures" analysis will be discussed in Chapter Six. Nevertheless, this problem is minor for most clinical studies where the number of patients is much larger than the maximal number of observations from each patient. In fact, if the patients' responses are more different over time than among them, the above method may end up to be more conservative. In general, the method presented here is simple and effective.

5.6 Far-from-average individual response profiles

An individual response profile is the response profile of a single patient over time. Because the responses of any individual patient are influenced by innumerable factors, it is fruitless, in general, to examine every individual response profile. It is those far-from-average individual response profiles that are most informative. Surprising discoveries are often made after scrutiny of far-from-average individuals.

Residuals are the measure of choice for identifying far-from-average individuals. A residual is the difference between data value and the mean. Residuals are ready for use once the means have been obtained from the analysis of variance. Given residuals, the formation of individual residual profiles is simply the linking of the residuals of every patient over the time points. The following charts depict the residual profiles from the previous analysis of the disease progression data:

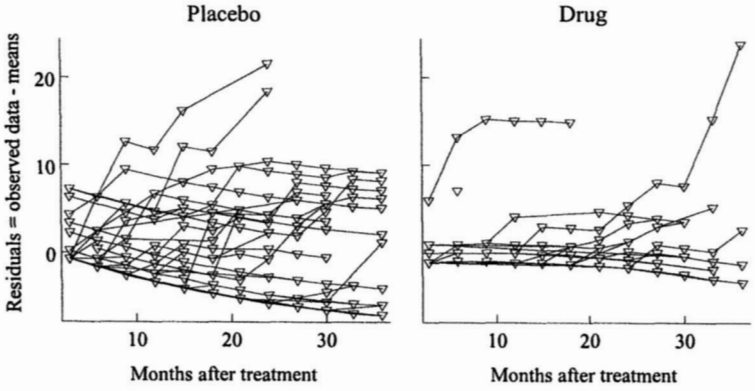


Figure 5.5 Deviation of individual response profiles from the means by treatment

Without treatment, the progression of the disease was quite heterogeneous and might not be fully characterized with a single average response curve. With treatment, however, the progression was fairly consistent and is well characterized by the means from the analysis of variance, except for only

two patients. This pattern of dispersion is what a great drug should demonstrate.

6

Nonparametric Analysis, Analysis on a Complex Scale, Analysis of Longitudinal Data, and Mixed Linear Models

Summary

The methods presented in this chapter are still in the framework of analysis of variance because the mean and standard error are still the main measures to characterize data. In essence, nonparametric analysis is the analysis of variance on transformed data, analysis on a complex scale compares functions of the means instead of the means themselves, analysis of longitudinal data is the analysis of variance with the standard errors derived from an average over the number of patients instead of observations, and finally, mixed linear models afford a means to compare far-from-average individual response profiles. Nonparametric analysis and analysis on a complex scale are discouraged due to the technical complexity and lack of scientific benefit. The analysis of longitudinal data presents an alternative approach to evaluate data with abundant within-patient observations. Mixed linear models are occasionally useful for selecting far-from-average individual response profiles.

6.1 An overview

The analytic techniques to be presented in this chapter are still under the broad umbrella of analysis of variance, because the data are still summarized with the mean and standard error, and the residual mean sum of squares is still the common numerator for all the standard errors. However, compared to the traditional analysis of variance methodology discussed in Chapter Four, these analytical techniques are much more complex; yet it is questionable, at least in my opinion, whether they are more scientifically advantageous in improving our understanding of data from clinical studies. Nevertheless, it is beneficial for researchers to be familiar with these techniques, because these techniques have been associated with numerous extravagant claims in statistical literature, and requests to use them are frequent from statistical authorities in the review of study proposals or reports.

6.2 Nonparametric analysis

The word, nonparametric, really means no involvement of mathematical distributions. The use of mathematical distribution is entirely technical and is discussed in detail in Chapter Ten. The requests for nonparametric analysis often come from statistical authorities for hypothetical reasons that are adduced against the use of certain mathematical distributions on certain types of data. The most common allegation is that the data are *not* normally distributed. For some researchers, more often than not, the real motivation behind nonparametric analysis is seeking a lucky p-value for making a claim. The most commonly performed "nonparametric" analyses are essentially the traditional analysis of variance on transformed data, although, in theory, permutation test, instead of the standard normal distribution or its equivalence, should be used to obtain the p-value for a nonparametric test. Examples of nonparametric analysis are the Wilcoxon, Mann-Whitney, and Kruskal-Wallis tests on ranks. Technically speaking, mathematical distributions can be used to expedite the computation in these nonparametric analyses to obtain "asymptotic p-values." In this sense, those nonparametric analyses are not truly nonparametric. I will not elaborate nonparametric statistics any further. Readers who are not familiar with nonparametric statistics may find it in many applied statistical textbooks.

Ranking is a popular protocol of data transformation for nonparametric analysis. Ranking preserves the order but ignores the magnitude of original observations. When the variation of data is large, ranking magnifies small values and dampens large values, which is thought to be advantageous by some who are interested in a reduced variation and is thought to be disastrous by others who are concerned with loss of information. Speaking in terms of p-values, the results of analysis of variance on original data and their ranks are not capriciously different; in terms of other measures, however, the results are generally not comparable.

Proposals for data transformation are numerous. The scientific justification is, however, seldom seen. The transformed data do not necessarily bear any information in the original observations, and the consequence of data transformation is not predictable in general. For scientific research, researchers should strongly discourage any data transformation for any alleged rationale having to do with a mere statistical technicality.

6.3 The analysis of variance on a complex scale

With the traditional analysis of variance, the means are compared directly. With the analysis of variance on a complex scale, directly compared are functions of the means. A scale is a function. A familiar function is logit for log odds:

$$\log\left(\frac{\text{mean}}{1-\text{mean}}\right),$$

which is commonly requested for the analysis of dichotomous categorical data. The logit transformations of 0.4 and 0.8, for instance, are –0.41 and 1.39. With the analysis of variance on the logit scale, it is –0.41 and 1.39 that are directly compared, as opposed to 0.4 and 0.8 with the traditional analysis of variance.

A computation tool for the analysis of variance on a complex scale is linear model on the scale of choice; for example,

g [mean(up)] = center + treatment + center-treatment interaction,

where g denotes the function. A linear model on an arbitrary scale is called generalized linear model. Two special cases of generalized linear models are logistic regression models and log-linear or Poisson regression models. The former is on the logit scale, and the latter is on the logarithmic scale. Of course, if we define the mean itself as being on the identity scale, the traditional linear models are also special cases of generalized linear models.

Much as the same distance may be measured on the metric or non-metric scale, whatever the scale of choice is does not have much impact on data evaluation in principle. In reality, however, a complex scale often imposes some restrictions on mathematical manipulations so that the actual results of analysis of variance on different scales can be slightly different. Perhaps the best way to study the effects of scales on the results of analysis of variance is by comparing the means and their standard errors from models on the scales of choice. The mean sums of squares from models on different scales are, however, not directly comparable.

The means are robust to change of scales as long as the explanatory variables are all categorical. The standard errors are less robust but generally agree from scale to scale. The following analyses assess the effects of center, treatment and center-treatment interaction on the identity, logarithmic, and logit scales:

$$\begin{matrix} \text{Mean(up)} \\ \text{Log[Mean(up)]} \\ \text{Logit[Mean(up)]} \end{matrix} = \text{center} + \text{treatment} + \text{center - treatment interaction.}$$

The standard errors of the means are converted to the identity scale with this formula:

$$\text{var}[g^{-1}(x\beta)] = \left(\frac{\partial g^{-1}(x\beta)}{\partial (x\beta)} \right)^2 \text{var}(x\beta)$$

See Appendix B for notations and details. The following table lists the means and their standard errors from the analyses:

		---------Means----------			----Standard Errors-----		
Center	Treatment	Identity	Log	Logit	Identity	Log	Logit
0001	Drug	0.1111	0.1111	0.1111	0.0373	0.0371	0.0372
	Placebo	0.0750	0.0750	0.0750	0.0500	0.0409	0.0418
0002	Drug	0.1000	0.1000	0.1000	0.0354	0.0334	0.0337
	Placebo	0.0250	0.0250	0.0250	0.0500	0.0236	0.0248
0003	Drug	0.1250	0.1250	0.1250	0.0646	0.0682	0.0678
**	***Placebo***	***0.0000***	***0.0000***	***0.0000***	***0.1582***	***0.0000***	***0.0000***
0004	Drug	0.0562	0.0562	0.0562	0.0335	0.0237	0.0245
	Placebo	0.0200	0.0200	0.0200	0.0448	0.0189	0.0199
0005	Drug	0.2162	0.2162	0.2162	0.0520	0.0722	0.0680
	Placebo	0.0625	0.0625	0.0625	0.0791	0.0591	0.0608
0006	Drug	0.0963	0.0963	0.0963	0.0272	0.0252	0.0255
	Placebo	0.1702	0.1702	0.1702	0.0462	0.0569	0.0551
0007	Drug	0.0896	0.0896	0.0896	0.0387	0.0346	0.0350
	Placebo	0.1053	0.1053	0.1053	0.0726	0.0703	0.0707
0008	Drug	0.2000	0.2000	0.2000	0.1415	0.1890	0.1797
	Placebo	0.1111	0.1111	0.1111	0.1055	0.1050	0.1052
0009	Drug	0.1638	0.1638	0.1638	0.0294	0.0355	0.0345
	Placebo	0.1311	0.1311	0.1311	0.0405	0.0438	0.0434
0010	***Drug***	***-.0000***	***0.0000***	***0.0000***	***0.1119***	***0.0000***	***0.0000***
0011	Drug	0.1284	0.1284	0.1284	0.0303	0.0324	0.0322
	Placebo	0.2167	0.2167	0.2167	0.0409	0.0568	0.0534
0012	Drug	0.2000	0.2000	0.2000	0.1001	0.1337	0.1271

Two noteworthy differences are highlighted in the table. One is that the mean of drug D in center 10 is 0.0000 on the log and logit scales, whereas it is –0.0000 on the identity scale. This is because negative values are not admissible by the definition of logarithm. This restriction on negative values is desirable for analyzing categorical data but presents a problem for data with negative values. The other difference is that the standard errors on the log and logit scales are all zero when the means are zero, and this is not happening on the identity scale. The cause of this difference is the functional association of mean and its standard error on the log and logit scales so that by mathematical definition, when the mean is zero, its standard error is always zero. This functional association between the means and their standard errors is one of many technical disadvantages with complex scales.

Note that this functional association between the means and their standard errors in the analysis of variance on a complex scale does not imply that the variances of the means are really some complex functions of the means as advocated in some statistical textbooks. The means and their variances are associated only in the sense that the means determine their variances. This is because the sufficient statistics for the variance of mean is the residuals, and once the mean has been determined from the

data, the residuals are simply the difference between the data and the mean. That functional association merely means that the variance is *parameterized* as a function of the mean. Parameterization is naming, and technically, it is a matter of convenience. As a result of using a complex scale, instead of a single Greek letter σ^2, a complex function of mean, such as $\sigma^2 f(\mu)$, is used to represent the variance, where μ denotes mean, and $f(\mu)$ denotes a mathematical function of the mean. If the mean is given and its variance is, say, 10, the whole matter of parameterization is to let $\sigma^2 f(\mu)$, not that simple Greek letter σ^2, to represent that variance of 10. It is exactly like naming a girl Nicholas rather than Katherine. Although this naming is bit exotic, it does not do any harm as long as the function contains an independent parameter other than the mean. Problem arises when the function contains only the mean, for example $f(\mu)$, and is used to represent the variance. This happens when the Poisson distribution is used to represent the frequencies of data, and the problem is referred to as "overdispersion or underdispersion." What happens is that once the mean is determined, the variance is restricted by function $f(\mu)$ to fully represent the variation of data as measured with the residuals. If the variance estimated with residuals is larger than that determined with function $f(\mu)$, overdispersion occurs; if smaller, underdispersion occurs. A solution to this problem is to introduce an element, ϕ, independent of the mean, μ, so that function $f(\mu,\phi)$ can be any value without being restricted by the value of μ. By definition, $f(\mu,\phi)$ and μ are functionally related, but they are independent in representing different measures. Since Katherine is a beautiful name, we may just use κ to replace that awkward $f(\mu,\phi)$. The Chinese proverb that bad names make difficult conversations literally speaks up the problem.

Complex scales do have some impact on the results of analysis of variance involving continuous explanatory variables. The following charts show the mean dose-response curves from the analysis specified with the following models on the scales of identity, complementary log-log, and logit,

$$\text{mean(up)}, \log[-\log(1 - \text{mean(up)})], \text{logit}[\text{mean(up)}] = \text{dose} + \text{treatment:}$$

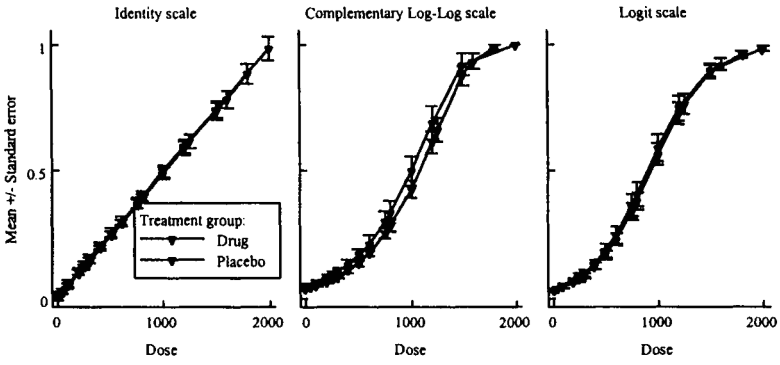


Figure 6.1 Mean response curves defined with three different scales

Another example is comparing the mean response profiles from the data of disease progression analyzed in Chapter Five on the identity and logarithmic scales:

$$\begin{matrix} \text{mean(freq)} \\ \text{log[mean(freq)]} \end{matrix} = \text{poly(time,2)} + \text{treatment} + \text{poly(time,2)} - \text{treatment interaction},$$

where poly (time, 2) denotes polynomial curves up to quadratic order.

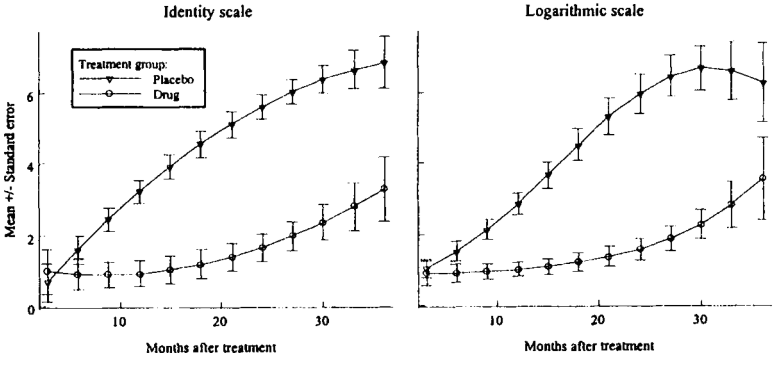


Figure 6.2 Mean recurrence over time by treatment on two different scales

The mean response curves are noticeably different upon change of scales, although they generally agree around the average follow-up month after treatment, the continuous explanatory variable. Unlike categorical variables whose effects are characterized by a group of means, the effects of continuous explanatory variables are characterized by the mean response curves, the shape of which has to be specified prior to the analysis. A scale defines a family of curves and resulting response curves cannot go beyond the family. For instance, the identity scale defines straight lines, and the logit scale defines sigmoidal curves. If the analysis uses a scale that defines straight lines, the resulting mean response profiles cannot be curly. In this sense of restriction, scales do matter in the analysis of variance involving continuous explanatory variables. However, as long as the scales of choice are not too restrictive, as they would be if they define, for instance, straight lines for data scattering in circle, the results of analyses on these scales should not be capriciously different.

In clinical research, scientific justification for complex scales hardly exists. The request for the analysis of variance on a complex scale often comes from statistical authorities who are, for hypothetical reasons, against the use of certain mathematical distributions under certain circumstances. Unless there is a reason for others, the identity scale should always be the scale of choice. When it is inevitable to perform the requested analyses under the pressure from authorities, researchers should not hesitate to proceed, knowing that, in general, the result of analysis of variance depends more on the choice of summary measures than on the choice of scales. Most of the time, change of scale results in 4 being compared to 8 while previously it was 1 being compared to 2.

6.4 Recent proposals for the analysis of longitudinal data

Data sequentially collected from a subject over time are called longitudinal data or repeated measures. Data from most clinical studies are longitudinal. The recent proposals for the analysis of longitudinal data are based on the thought that longitudinal data are correlated. In statistical literature, correlation, when spoken by different people, could mean completely different things. For longitudinal data, correlation could mean either that a patient's response at a moment is closely related to this

patient's previous responses, or that a series of responses come from a Chinese patient. The consensus is that the responses from each patient should be viewed as a unit, and it is the number of patients, not the number of observations, that should be used to measure the strength of evidence and compute the standard errors.

The recently proposed methods are generally referred to as "repeated measures" or longitudinal data analysis. The methods are still in the broad scope of analysis of variance, but the standard errors are defined differently. While the standard errors in the traditional analysis of variance are based on an average over the total number of observations, even though they all come from a single patient, the standard errors in a repeated measures analysis are derived from an average over the number of patients, no matter how many observations are made from each patient. Thus, when a large number of observations are made from each patient and the variations of within-patient observations are smaller than the variations of between-patient observations, the newly defined standard errors tend to be larger than their traditional counterparts.

Technically, the residual variations due to the uncontrolled factors are represented with a matrix, known as the residual matrix, in linear models for repeated measures analysis, while they are represented with a scalar in linear models for the traditional analysis of variance. Linear models for repeated measures analyses are often referred to as repeated measures models or generalized estimating equations (GEEs) for scales other than identity. The technical complexity for fitting repeated measures models can be formidable, but the result is not always satisfactory. This is mainly because the residual matrix cannot be reliably estimated when missing data are abundant or the number of within-patient observations is larger than the number of patients. The matrix then has to take some arbitrary structure. To minimize the impact of this arbitrariness on the result of analysis, the current solution is to derive the standard errors directly from the residuals. In statistical literature, such directly estimated standard errors have many names like "empirical estimators," "sandwich estimators," or "robust estimators," just to name a few.

Compared to its traditional counterpart, the results from an appropriate repeated measures analysis of variance are not very different for most clinical studies where the number of patients is much greater than the

maximal number of observations from each patient. Suppose about 900 patients are assigned to four treatment groups and are evaluated over 16 visits. The following models specify the effects of interest.

responses = treatment + visit + treatment-visit interaction + residuals
(scalar, unstructured matrix, autoregressive matrix).

The model with a scalar is a traditional linear model, and the models with matrices are repeated measures models. The unstructured and autoregressive matrices are two of many arbitrary structures that a matrix can take. The following table compares the means and their standard errors for treatment A:

Table 6.1 Comparison of Means and Standard Errors from Two Models

		Traditional Linear		Repeated Measures Models			
		----Model---		--unstruct--		--autoreg---	
Treatment	Visit	Mean	Stderr	Mean	Stderr	Mean	Stderr
Drug A	0	1.62	0.0620	1.62	0.0582	1.61	0.0582
	1	1.44	0.0618	1.44	0.0616	1.44	0.0616
	2	1.35	0.0622	1.36	0.0603	1.36	0.0603
	3	1.32	0.0622	1.33	0.0607	1.32	0.0607
	4	1.24	0.0622	1.25	0.0602	1.25	0.0602
	5	1.30	0.0625	1.32	0.0615	1.31	0.0617
	6	1.26	0.0627	1.28	0.0596	1.27	0.0597
	7	1.28	0.0628	1.30	0.0608	1.29	0.0610
	8	1.22	0.0630	1.24	0.0629	1.23	0.0631
	9	1.20	0.0630	1.22	0.0624	1.21	0.0626
	10	1.21	0.0631	1.22	0.0654	1.21	0.0658
	11	1.16	0.0634	1.16	0.0600	1.16	0.0604
	12	1.15	0.0634	1.17	0.0619	1.16	0.0626
	13	1.10	0.0636	1.12	0.0605	1.11	0.0614
	14	1.07	0.0640	1.09	0.0608	1.08	0.0617
	15	1.07	0.0648	1.08	0.0626	1.07	0.0635
	16	0.96	0.0754	0.98	0.0663	0.97	0.0674

unstruct: Unstructured
autoreg: The first-order autoregressive structure

While the means are almost identical, their standard errors agree up to the second decimal point. In this analysis, the means are actually computed in each treatment group at each visit, and they are linked over visits to form the mean response profiles. This is feasible because there are sufficient observations at each visit. With the same data, the mean response profiles

are specified as polynomial curves and compared with the following models:

responses = treatment + poly(days, 2) + treatment-poly(days, 2) interaction + residuals (scalar, unstructured matrix, autoregressive matrix).

The results are summarized in the following graph with comparison to the mean response profiles directly computed from the data without using a linear model:

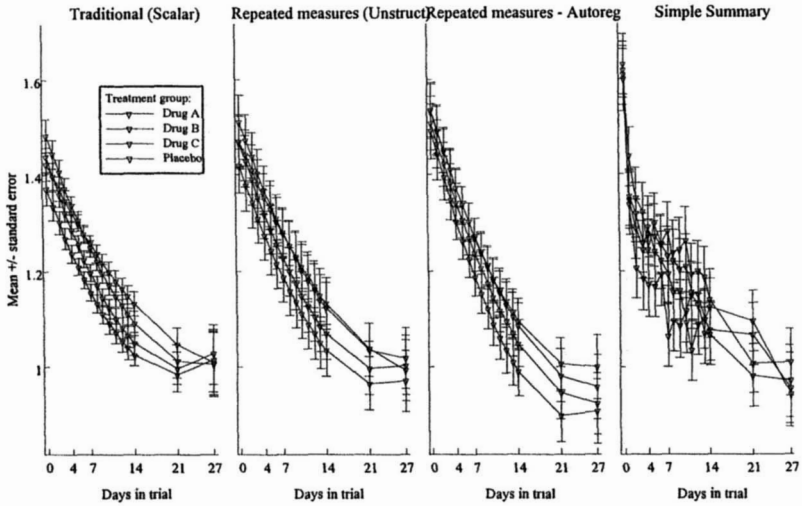


Figure 6.3 Comparison of means and standard errors from linear models with different parameters for residual variations

Once again, the analyses with linear models give rise to similar results, and they seem to agree to the result of simple summary without a linear model.

The repeated measures type of analysis is perhaps most useful when the number of observations from each patient is much larger than the total number of patients, and there is a deep concern with the representability of data to the patient population. The following analysis presents the profiles of success rate from a trial in which 18 patients were assigned to three

treatment groups, each patient was sequentially tested 180 times, and the responses are dichotomous, with 1 for success and 0 otherwise. The effects of interest are specified with the models:

responses = poly(seq. No., 2) + treatment + poly(seq. No., 2)-treatment interaction + residuals (scalar, autoregressive matrix),

where the model with a scalar represents the traditional analysis of variance, and the model with an autoregressive matrix represents a repeated measures analysis of variance. The means and their standard errors are presented in the following two charts:

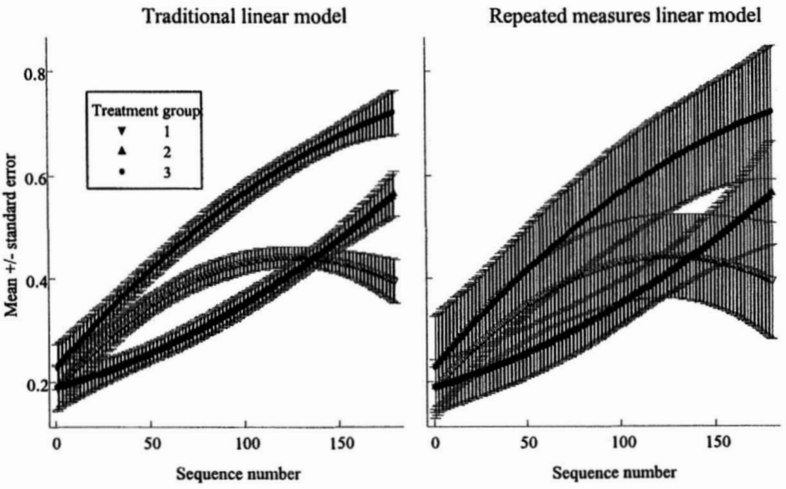


Figure 6.4 Comparison of means and standard errors from linear models using scalar and matrix for residual variations

The means from both analyses are almost identical. However, the standard errors from the repeated measures model are 2 or 3 times as large as those from the traditional linear model. The large standard errors reflect the fact that the means are based on 18 patients rather than 18 × 180 = 3240 observations.

6.5 Analysis of longitudinal data with mixed linear models

Mixed linear models contain both fixed and random effects terms. The model was proposed by Charles R. Hendersen, who was primarily interested in the deviations from the means than the means themselves. His purpose was selecting extraordinary, not average, animals for breeding. The deviations were what he meant by random, and the means were what he meant by fixed. All the linear models discussed thus far are for the computation of means, and they are collectively called fixed effects models. The name of random effects has indeed generated a good deal of confusion. Some statistical authors even link the name to the process of randomization or sampling, saying that mixed models especially suit longitudinal data because the patients are randomly samples and, therefore, should have some random effects. A better name that appeals to me is individual effects, reflecting that random effects truly represent the deviations of individual patients from their averages.

When the primary interest is means, mixed linear models are *not* any better than fixed effects linear models. For longitudinal data, mixed linear models afford at most a means to define a matrix to represent the residual variations due to the uncontrolled factors. But when the residual matrix is estimated directly from residuals, whatever structure of choice is inconsequential. As an example, the sequential testing data on 18 patients are analyzed with the mixed linear model,

Fixed effects:	responses = poly(seq. No., 2) + treatment
	+ poly(seq. No., 2)-treatment interaction
Random effects:	+ intercept + poly(seq. No., 2) for every patient
Residuals:	+ residuals

The result is graphically compared to that from a traditional linear model with identical fixed effects,

responses = poly(seq. No., 2) + treatment+poly(seq. No., 2)-treatment interaction + residuals:

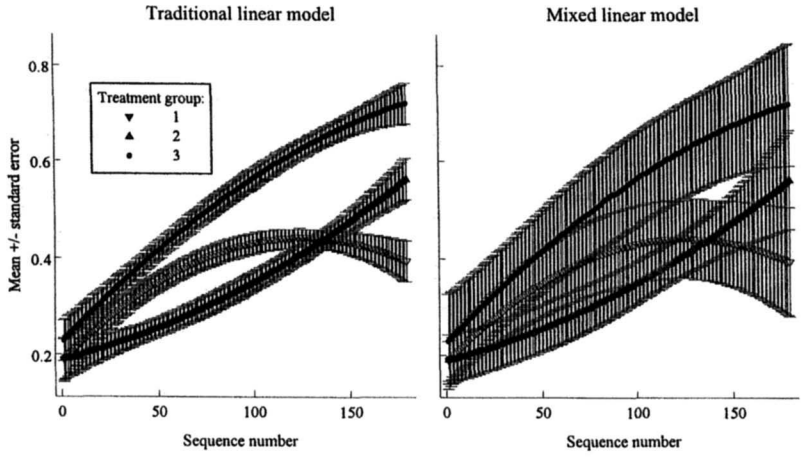


Figure 6.5 Comparison of means and standard errors from fixed effects and mixed effects linear models

Compared to its fixed effects model counterpart, the mixed effects model first computes the mean quadratic polynomial response curves and then the deviations between the mean curves and the quadratic polynomial response curves for each patient:

$$\text{intercept} + \text{poly}(\text{seq. No.}, 2).$$

It is not surprising that the means from these two models are almost identical. Indeed, if any difference is noticed after addition of random terms in a model, one should be very skeptical of the validity of computation. Comparing Figures 6.4 and 6.5, the standard errors from the mixed linear model are not different from those from the previous analysis with the repeated measures model. As a matter of fact, as long as the residual variation matrix is directly estimated from the residuals, the standard errors hardly change at all, no matter what random effects are specified in a mixed linear model.

Mixed linear models are useful only when the primary interest is far-from-average individual response profiles. Compared to passive linking of residuals, a mixed linear model affords a means to specify a pattern for

each patient. The following graph presents the residuals from the traditional linear model and the random or individual effects from the mixed linear model for the sequential testing data, where the random or individual effects are specified as,

intercept + poly(seq. No., 2), for every patient,

with poly (seq. No., 2) denoting quadratic polynomial curves:

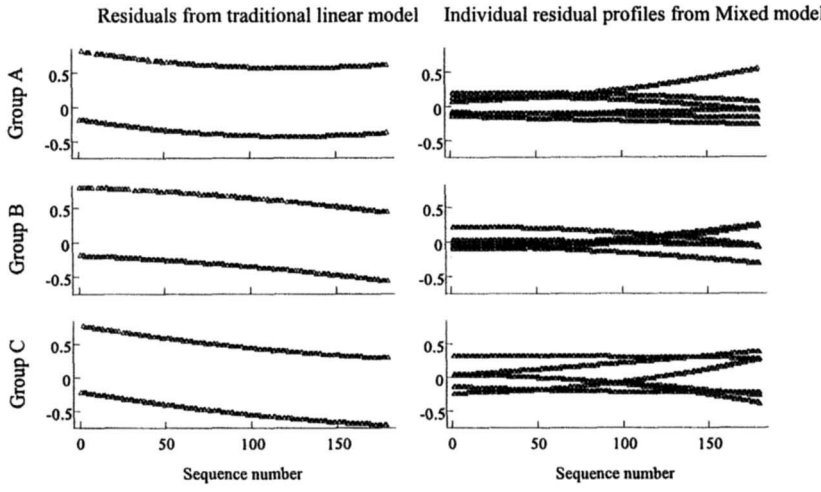


Figure 6.6 Comparison of individual residual profiles from fixed effects and mixed effects models

Because the responses are dichotomous (0 or 1), the residuals clump together, and the resulting graph provides little information for figuring out far-from-average individual response profiles. With mixed linear model, each patient's response profile is represented with a polynomial curve, and this curve is compared to the mean response curves. There are six curves in each treatment group, each curve representing the deviation of that individual's response profile from the mean response profile.

This Page Intentionally Left Blank

7
Survival Analysis

Summary

The constant flow of time makes survival observations from patients lost in follow-up meaningful. Comparisons of survival time values themselves are not quite meaningful when a considerable number of observations are from the patients lost in follow-up. An appropriate measure for summarizing survival observations is time-specific death or survival rate, which is the ratio of deaths or lives at a specific time point over the number of patients at risk of death. Comparisons of survival information may be made with life tables or Kaplan-Meier plots. The former is cross tabulation of time-specific death rates, and the latter is cross display of cumulative survival rates. The analysis of variance on survival data may be carried out with linear models as described in Chapter Four or Cox's regression models.

7.1 Survival data and summary measures

The constant flow of time makes survival observations from patients lost in follow-up meaningful. For instance, that a patient died at age of 45 and that a patient was last found alive at age of 45 are equally informative. The first patient survived for 45 years, and the second patient survived for

at least 45 years. For non-survival data, however, loss of follow-up means complete loss of information. For instance, that a patient's blood sugar was not known at age of 45 due to loss of follow-up means a complete lack of information on that patient's blood sugar at that time. Because time values themselves are not indicative of death and loss of follow-up, the recording of survival information requires at least two numbers, one for the time and the other for the event that associates with the time. If we choose 1 to indicate death and 0 to indicate loss of follow-up, survival information can be recorded as (time, death = 1 or 0). In the statistical literature, loss of follow-up is also known as censoring, and the associated time values are often referred to as censored survival data.

Because time values carry only partial survival information, regular measures focusing on time values, such as the mean and standard deviation, are not quite meaningful unless none of the time values is associated with loss of follow-up, or in other words, none is censored. An appropriate measure for survival data is time-specific death rate or survival rate:

$$\text{time-specific death rate} = \frac{\text{number of patients died at a time}}{\text{number of observable patients right before the time}},$$

$$\text{time-specific survival rate} = \frac{\text{number of patients alive at a time}}{\text{number of observable patients right before the time}}.$$

These two rates are complementary: death rate = 1 – survival rate. The number of deaths or lives contributes to the numerator. The denominator is the number of patients at risk of death. Those who died or are not observable due to loss of follow-up before the time are not counted toward the denominator. Therefore, the number of losses in follow-up contributes to the value of denominator.

When the causes of death are known, their deadly force may be measured with cause and time-specific death or survival rate:

$$\text{cause and time-specific death rate} = \frac{\text{number of patients died of the cause at a time}}{\text{number of observable patients right before the time}},$$

$$\text{cause and time-specific survival rate} = \frac{\text{number of patients not died of the cause at a time}}{\text{number of observable patients right before the time}}.$$

Here "survival" means no death from the cause, not the literal sense of being alive.

7.2 Cross tabulation of death and projected survival rates

Tables showing death rates and their derivatives are known as life tables. The essential steps in construction of a life table are grouping time values into intervals and then counting the number of deaths and the number of patients at risk:

Table 7.1 Life Table

Survival time (+: censored)	Interval	Death	Other losses	Patients at risk	Death rate	Projected survival rate
8, 8, 9	0 – 10	3	0	14	3/14	78.57 %
10, 12+, 12, 12, 13, 15	10 – 20	5	1	10	5/10	39.29 %
20+, 20, 24+	20 – 30	1	2	4	1/4	29.46 %
30+, 34	Beyond	1	1	2	1/2	14.73 %

In this table, patients at risk is the number of observable patients at the beginning of an interval, and projected survival rate is the cumulative product of survival rates from the first interval; for instance,

$$\text{projected survivival rate beyond 20} = (1-\frac{3}{14}) \times (1-\frac{5}{10}) \times 100\ \% = 39.29\%.$$

The projected survival rate is an estimate of the percentage of patients who may survive longer than a specific time interval. The logical basis of this multiplication of time-specific survival rates is that if a patient is found alive at a time point, this patient must have survived all the way through to the point. The projected survival rate is a projection of time-specific survival information rather than a simple ratio of lives and patients at risk, such as 42% = 6/14 at the end of 20. The difference is that the projected survival rate is adjusted for the number of losses in follow-up, while the simple ratio is not.

The death rate is time-specific, while the projected survival rate is both time-specific and cumulative; both are the commonly used measures for summarizing survival data. However, these two measures must be interpreted together with other two important pieces of information: the number of patients at risk and the number of patients lost in follow-up. The number of patients at risk determines the reliability and robustness of the estimated death rate against the background of the uncontrolled factors. Any rate based on a small number of patients at risk is not reliable because the contribution from each patient is too large to warrant an adequate control of confounding from the uncontrolled factors. When two patients are at risk, for instance, the death rate can be different by 50% upon the fate of a single patient. Loss of follow-up, on the other hand, potentially confounds with death rates, especially when the underlying causes of death have direct impact on the number of patients lost in follow-up. Thus, a complete summarization of survival information requires four measures: death rate, projected survival rate, loss of follow-up, and patients at risk.

Cross tabulation of those four measures expedites the comparison of survival information across groups. The following table is an example where the time-specific death rates, projected survival rates, loss of follow-up, and patients at risk are denoted by death, survival, LoF, and at risk:

Table 7.2 Summary of Survival Data by Treatment

Group	Measure	0 - 10	10 - 20	20 - 30	30 – 40	40 - 50	50 +
Treatment	**Death**	9.1%	20.0%	14.3%	40.0%	33.3%	0.0%
	Survival	90.9%	72.7%	62.3%	37.4%	24.9%	24.9%
	LoF	0	1	1	0	1	1
	At risk	11	10	7	5	3	1
Placebo	**Death**	33.3%	12.5%	33.3%	50.0%	100%	
	Survival	66.7%	58.3%	38.9%	19.4%	0.0%	
	LoF	0	1	0	0	0	
	At risk	12	8	6	4	2	

The projected survival rates and patients at risk are graphically presented:

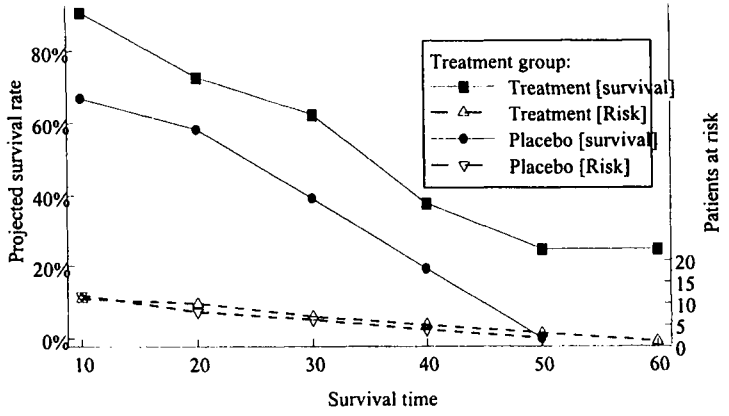


Figure 7.1 Projected survival rates and patients at risk by treatment

7.3 Kaplan-Meier plot of projected survival rates

The result of a life table depends, more or less, on the grouping of survival time values. When the number of observations is small, grouping could produce artifacts. An alternative is to calculate projected survival rates at the moments of death, instead of in arbitrarily defined intervals, and a graphical presentation of the resulting projected survival rates is known as the Kaplan-Meier plot. The projected survival rate at the moment of death, t, is the cumulative product of survival rates at times of death before t:

projected survival rate (t) =
product (1 – death rates at times of death before t).

The projected survival rates at all times of death are referred to as the product-limit or Kaplan-Meier estimate of survival function. The death rate at any moment is not very informative as a global assessment of survivorship, not only because only one death occurs at a time mostly, but also because deaths do not occur at the same time so that comparisons of death rates can not be made in a common time frame. On the other hand, projected survival rates are cumulative and, therefore, much smoother.

They generally measure a process of survivorship within a period of time rather than sporadic events at specific time points.

The following is a Kaplan-Meier plot of projected survival rates by stratum and treatment:

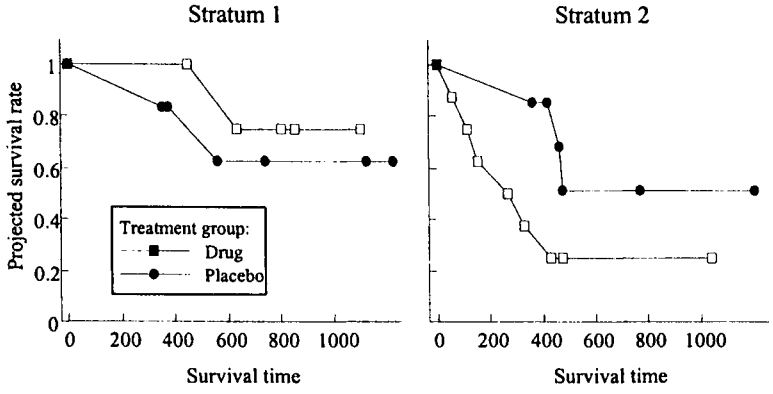


Figure 7.2 Kaplan-Meier plot of projected survival rates by stratum and treatment

When the curves become flat, the survival times are when the patients were last seen, meaning that the patients survived beyond the end of the study.

In general, survival curves are not quite reliable when they reach the end, because only few patients remain under observation. The number of patients at risk is the single most important indicator for the reliability and robustness of any estimate. It is questionable whether deviation measures such as the standard deviation have any meaning for a single survival rate. For a single time-specific death or survival rate, a measure of reliability is the reciprocal of patients at risk,

$$\text{contribution to time - specific death or survival rate} = \frac{1}{\text{patients at risk}},$$

which measures the contribution of each individual patient to the rate. Suppose 3 deaths occur at a time when 5 patients are at risk. The death rate is 3/5 = 60%. If the fate of a patient had changed due to the effects of the uncontrolled factors, the death rate would have been either 4/5 = 80% or 2/5 = 40%. Thus, the fate of a patient could affect death rate by 1/5 = 20%. Since the projected survival rate is the cumulative product of time-specific survival rates, from

$$\text{projected survival rate at time t} = \text{projected survival rate at time (t - 1)} \times (\text{survival rate at t} \pm \frac{1}{\text{patients at risk}}),$$

we may measure the impact of each patient at a time point on projected survival rate by

$$\text{impact on projected survival rate at time t} = \frac{\text{projected survival rate at the previous time (t - 1)}}{\text{patients at risk right before time t}},$$

which is the difference that the fate of a single patient could make on the projected survival rate at time t. With this measure of impact, the projected survival rates are replotted in the following graph:

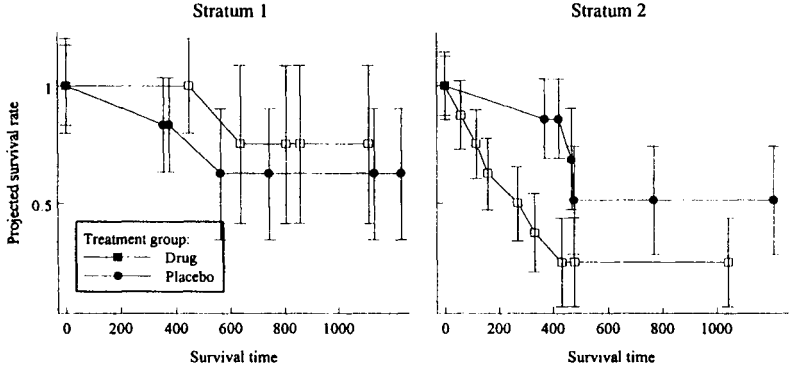


Figure 7.3 Kaplan-Meier plot of projected survival rates with measure of impact from a single patient

The vertical line extending from the survival curves at a point indicates how much the curve could vary upon the fate of a single patient at that point.

7.4 The analysis of variance on survival data

The analysis of variance technique in Chapter Four may be directly used to compare time-specific death or survival rates. However, the survival data need to be specially arranged in such a way that, at any time of death, the dead patient or patients, and patients at risk can be explicitly identified. The following table illustrates the data structure ready for the analysis of variance with traditional linear models:

Table 7.3 Data Rearranged for a Direct Count of Deaths with Linear Model

Original Data Structure			
Variables:	**Time**	**Death**	**x**
One record for a patient	10	1	1
	10	1	2
	20	0	3
	20	1	4
	30	0	5
	40	1	6
Death = 0 means loss of follow-up.			
Rearranged Data Structure			
New variables:	**Time**	**Death**	**x**
At time 10:	10	1	1
2 deaths	10	1	2
6 at risk	10	0	3
	10	0	4
	10	0	5
	10	0	6
At time 20:	20	0	3
1 death	20	1	4
4 at risk	20	0	5
	20	0	6
1 death / 1 at risk	40	1	6
Death = 0 means alive at the time.			

In the original data set, deaths and patients at risk are implicitly indicated by the time values of death and loss of follow-up. In the rearranged data set, the information of all observable patients at any specific time of death is explicitly presented. The records are indexed by the time of death. The time of loss in follow-up contributes only to patients at risk. "x" is a explanatory variable that may change its value over time. The survival time values are used as a chronological marker to figure out whether or not a patient has died at a specific time and the values of time-dependent explanatory variables.

With explicit death and risk information, the analysis of variance for comparing time-specific death or survival rates is straightforward. Suppose about 30 patients from Africa and Asia are randomly assigned to drug and placebo, and the purpose of the study is to compare the effects of drug on patients' survival. The effects of interest are specified in the following linear model:

$$\text{deaths} = \text{continent} + \text{time} + \text{treatment} + \text{time-treatment interaction} + \text{residuals},$$

where continent represents the effects of patients' origin on death rates, time represents the overall effects of time on death rates, treatment represents the effects of treatment on the combined death rates across time, and time-treatment interaction represents the effects of treatment on time-specific death rates. The death rates are compared on both the identity and logarithmic scales. The logarithmic scale prevents negative means. The results are summarized in Figure 7.4 on the next page.

Because only few patients remained in the study after 400 days, the claimed effects of time-treatment interaction might have been confounded with the effects of the uncontrolled factors. Without claiming the interaction effects, the treatment effects on the combined death rates are compared with the model,

$$\text{deaths} = \text{continent} + \text{time} + \text{treatment} + \text{residuals},$$

and the results are summarized in Figure 7.5. The results of analysis may also be presented with the survival curves estimated from the means:

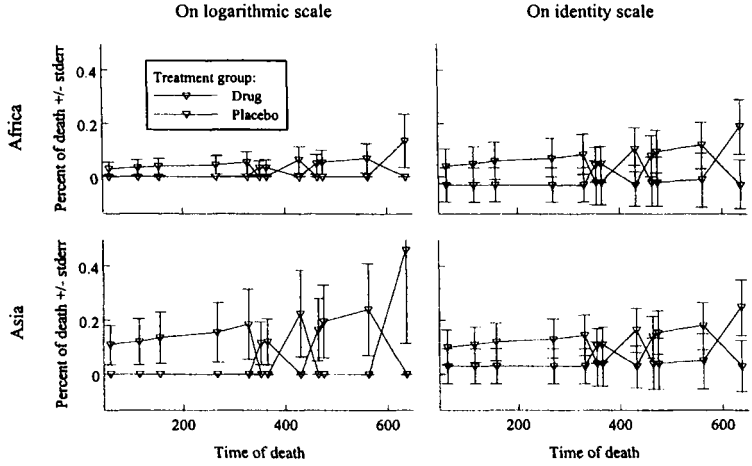


Figure 7.4 Time-specific death rates by continent and treatment from linear models on different scales

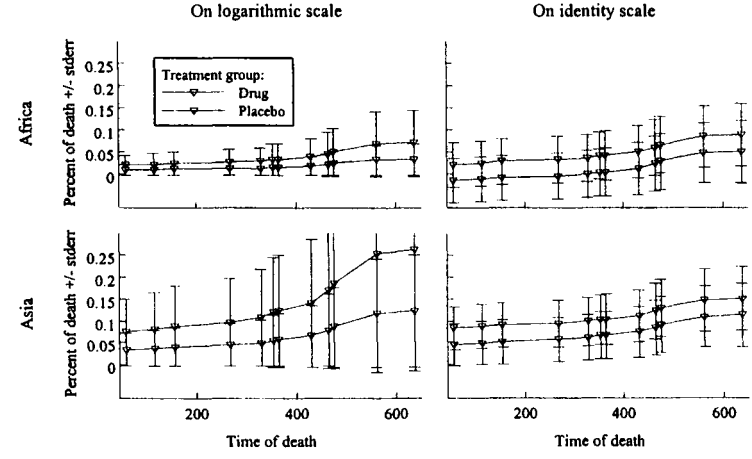


Figure 7.5 Time-specific death rates by continent and treatment from linear models on different scales under the assumption of equal survivorship

projected survival rate at time t = prod (1-mean)
at all death points up to t.

The survival curves are plotted in Figure 7.6. It appears that the patients on drug have a lower survival rate than the patients on placebo although the difference may not be very significant.

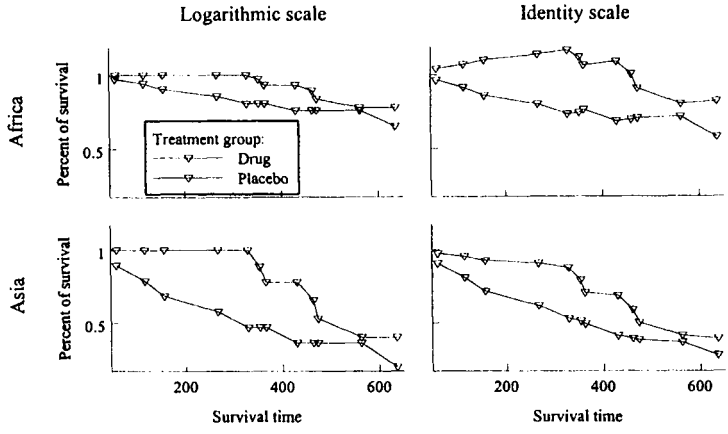


Figure 7.6 Projected survival rates based on the means from analysis of variance

7.5 Analysis of variance with proportional hazard models

Comparisons of time-specific death rates can also be made with a model proposed by David R. Cox, known as the proportional hazard model or Cox's regression model. Hazard is a synonym of death rate. The proportional hazard model is most conveniently specified at times of death; for instance,

death rate at time t $\propto$ *exp* (baseline + insulin dose at time t + treatment),

where *exp* denotes the exponential operator. The main purpose of operating on the exponential scale is to prevent negative values. The effects of interest are baseline, treatment and insulin dose. The effects of baseline and treatment are not time-dependent, and thus, they represent the

overall effects across all time points. The effect of insulin is time-dependent because the dose of insulin changes over time.

The basic operation in fitting a proportional hazard model is to count, at each time of death, the numbers of deaths and patients at risk, and to set the current values of explanatory variables. Time-dependent variables take different values at different time. This operation is identical to the rearrangement of data structure in section 7.4. Both utilize survival time values as a chronological marker to figure out the number of deaths, the number of patients at risk, and the values of time-dependent explanatory variables. In Cox's model, the death rate at each time point is actually represented by a ratio with unknown parameters,

$$\text{death rate at time t: } \frac{exp\text{ (effects of interest) for patients who died at time t}}{\text{sum of } exp\text{ (effects of interest) for patients at risk before t}}.$$

Numerical computational techniques are used to find a set of parameter values that maximize the joint product of these death rate representatives at all times of death.

Because of this peculiar parameterization, the direct estimate from Cox's model is risk ratio, not death rate. Risk ratio is the ratio of predicted deaths under different conditions specified with explanatory variables. The risk ratio between treatment groups, for instance, is obtained by

$$\frac{\exp\text{ (baseline + insulin + treatment = 1)}}{\exp\text{ (baseline + insulin + treatment = 0)}} = \exp\text{ (treatment = 1)},$$

where only the treatment variable changes values, and the others are held constant. The effects of time on death rates are difficult to evaluate directly with Cox's model. In order to evaluate the effects of time on death rates with Cox's model, the time variable has to be re-scaled, grouped, or associated with a time-dependent variable. The purpose is to gather sufficient number of deaths in each time interval so that the parameters can be reliably estimated.

On comparable measures, the results from a traditional linear model and the corresponding Cox's model are not capriciously different. The following analyses of survival data from an oncology trial are an example,

where the effects of recurrence, number and size of the tumor at diagnosis, and treatment on patients' survival are evaluated. The following table presents the results from the traditional linear model,

deaths = recurrence + number + size + time + treatment + residuals

and the Cox's models,

death at time t = *exp* (number + size + treatment),
stratified by recurrence,

and

death at time t = exp (recurrence + number + size + treatment):

Table 7.4 Comparison of Results Out of Linear and Cox Models

	Linear Model		Cox's Model (1)		Cox's Model (2)	
Source	**Risk Ratio**	**P-value**	**Risk Ratio**	**P-value**	**Risk Ratio**	**P-value**
Number		0.2673	0.95	0.4647	0.95	0.4423
Size		0.0001	1.23	0.0001	1.25	0.0001
Treatment	*0.64	0.0035	0.56	0.0039	0.55	0.0029

- * ratio of the least squares means for treatment, 0.0324/0.0504.
- Cox model (1): stratified by recurrence.
- Cox model (2): recurrence is an explanatory variable.

In these models, recurrence, treatment, and time are categorical variables, and number and size are continuous variables. The analyses suggest significant treatment effects over time, and the treatment appears to reduce the chance of death.

This Page Intentionally Left Blank

8
The Design of Clinical Study Programs

Summary

Control of confounding at the stage of data acquisition is the main focus. The idea is setting up comparable treatment groups so that any difference among the groups can be logically and reliably attributed to the effects of treatment, not the effects of other factors. Basic operations include setting up study groups, randomization in patient assignment to treatment, the use of control, blinding in evaluation of treatment effects, and finally, stratification to improve the precision of estimates. The determination of sample size is another main topic of this chapter. Statistical power calculation is criticized for its lack of logical basis and determination of sample size with non-observable measures. The criteria of sensitivity, stability, and precision are proposed to set the minimal number of patients.

8.1 An overview

Like any other scrupulous commercial business, clinical studies are an investment for useful results, and a clinical study program is a complex business process. A scientific and ethical standard is set forth in an international documentation, known as the Good Clinical Practice (GCP)

guidelines, which will be the topic of Chapter Eleven. The characteristic of clinical studies is the constant presence of innumerable known and unknown factors that potentially confound with the factors under study. While measurement of confounding effects has been the main focus in statistical analysis of available data, the main concern in this chapter is techniques for the control of confounding at the stage of data acquisition, so that the results derived from the data are valid and reliable.

Conclusion on treatment is valid if the difference between treatment groups can be logically attributed to the effects of treatment, not to the effects of other factors; conclusion on treatment is reliable if it is robust against the effects of the uncontrolled factors and can be consistently demonstrated in a series of studies. It generally requires an even distribution of all possible confounding factors among treatment groups to establish logical validity, and it generally requires adequate quantity of observations in treatment groups to ensure reliability.

Because of the complexity of human subjects and limitation of resources, most clinical study programs consist of a series of studies, each of which is designed to address a specific question. Each study is designed to be simple, with few groups under comparison and limited stratification. In general, a series of studies are sequentially carried out, and the study program is constantly adjusted, based upon the knowledge available from the completed studies. Eventually, the entire series are combined for a comprehensive account of the study topic.

8.2 Parallel and crossover setups

The foundation of clinical study is setting up groups for comparison. Parallel and crossover setups are the basic. In a parallel setup, each patient is assigned to only one treatment. If a patient receives treatment A, for instance, this patient will not receive treatment B. In a crossover setup, each patient is assigned to multiple treatments in a sequence. For instance, a patient may receive treatment A and then treatment B, and another patient may receive treatment B first and then treatment A.

8.2.1 Parallel setups

A parallel setup is simply a number of clearly defined groups to which patients are assigned in a mutually exclusive manner, so that the data can

be directly compared among the groups. It is straightforward to set up studies like this; for example, an investigational drug at 10 and 20 mg tid is compared to placebo and active controls:

Drug 10 mg tid	Drug 20 mg tid	Placebo control	Active control

where each cell represents a treatment group. In this setup, each treatment acts alone, and comparison among treatment groups is straightforward.

It is slightly complicated to explore the joint effects of multiple drugs. A special grouping of treatments, known as the factorial structure, due to Ronald A. Fisher, is very effective. The factorial structure for a placebo-controlled, two-drug study is best presented with a table:

	Drug A	Placebo
Drug B	**A + B**	**B**
Placebo	**A**	**P**

where each cell represents one of four possible combinations of the drug and placebo. If we rearrange this table of four cells into

Placebo	**Drug A**	**Drug B**	**Drugs A+B**

we will see that this is still a parallel setup, in which patients are allocated to four independent treatment groups. The difference from the previous unstructured parallel setup is that the combinations of treatments have an intrinsic structure. This intrinsic factorial structure allows for the effects of drugs A and B being evaluated both separately and in combination. The factorial setup for evaluating the joint effects of three drugs is shown as follows:

	Drug C		Placebo	
	Drug A	Placebo	Drug A	Placebo
Drug B	**A + B + C**	**B + C**	**A + B**	**B**
Placebo	**A + C**	**C**	**A**	**P**

This table may be viewed as a replicate of the 2x2 setup for A and B at C and placebo. If 30 patients are recruited to each of the eight groups,

Placebo	Drug A	Drug B	Drug C	Drugs A + B	Drugs A + C	Drugs B + C	Drugs A + B + C

this may still be a manageable study.

Simple factorial setup becomes inefficient to evaluate four or more drugs simultaneously. The number of groups shoots up exponentially in the order of 2^n, where n denotes the number of drugs. A setup with a huge number of groups, $2^4 = 16$ for 4 drugs for instance, creates insurmountable management problems that may result in slow patient recruitment and poor data quality. Furthermore, resources may be wasted on many drug combinations whose effects are likely to be undesirable.

It is more efficient to adopt a stepwise approach to study multiple drugs simultaneously. Suppose after the three-drug trial, we find that the B-C combination is the best, and we would like to add drug D. Instead of setting up a 4-drug factorial, 16-group trial, it would be cost-effective to setup up a 4-group trial,

Placebo	**Drug D**	**Drug B + C**	**Drugs B + C + D**

which may be viewed as another 2x2 factorial setup,

	Drugs B + C	Placebo
Drug D	**Drugs B + C + D**	**Drug D**
Placebo	**Drugs B + C**	**Placebo**

This setup allows us to evaluate the effects of drug D alone and together with the current best combination. It is possible, of course, that A-D is in fact the best combination, which clearly has been missed by this path of search. To search in a broader scope, one may add another two groups so that the setup becomes

Placebo	**D**	**B + C**	**A + D**	**B + D**	**D + B + C**

Another option is to run a separate trial with the following setup,

Placebo	**A + D**	**B + D**

and the results from the two studies are pooled together to find out the best combination.

In setting up a series of trials, it is extremely important to have a control group in each trial in recognition that the same drug may show completely different effects from trial to trial, due to the uncontrolled factors. The comparison of placebo groups across trials is an objective evaluation of the effects of trials. If the placebo groups are similar, other groups may be directly comparable across trials. If the opposite is true, directly comparable are only the within-trial contrasts with placebo, such as A – placebo in one trial and B – placebo in the other. Without adequate control groups, it is difficult to combine results from a series of trials. Furthermore, if the purpose is for comparison, the placebo group is as important as the groups receiving active treatments, for comparisons to placebo cannot be made precisely until precise information is available in both treatment and placebo groups. One must think about the purpose very carefully when planning uneven allocation of patients to placebo and active treatments.

8.2.2 Crossover setups

A crossover setup is a number of treatment sequences, with each patient assigned in a mutually exclusive manner to one of those sequences. Once assigned to a sequence, the patient receives multiple treatments in the designated order of that sequence. Take a 2 x 2 crossover setup as an example. Patients are assigned to treatment sequence A $\Rightarrow$ B or B $\Rightarrow$ A. Depending on the sequence, a patient receives treatment A or B for a period of time and, after a washout period, the alternate treatment for another period of time. The following table represents the setup:

	Period 1	Washout	Period 2
Sequence A$\Rightarrow$B	**Treatment A**	no treatment	**Treatment B**
Sequence B$\Rightarrow$A	**Treatment B**	no treatment	**Treatment A**

Notice that it is the sequences of treatment, not any particular treatment, that the patients are assigned to. This is important at the time of randomization.

The real motivation for crossover setup is reducing cost by generating large quantity of data from very few patients. The claim that crossover

trials improve precision by making within-patient comparisons is questionable in clinical research practice. The difference from time to time for a single patient can be just as much as the difference from patient to patient can be.

The ideal condition for crossover setup is when the patients are identical before receiving any treatment in any period. If this condition is reasonably true, there is no need to balance treatments in each period, and the use of multiple sequences is only for the purpose of blinding. For instance, instead of using four sequences to compare the effects of drugs A, B, C and D,

	Period 1	**Period 2**	**Period 3**	**Period 4**
Seq. 1	A	B	C	D
Seq. 2	B	C	D	A
Seq. 3	C	D	A	B
Seq. 4	D	A	B	C

only two of the four sequences may be selected so that neither patients nor evaluators are able to figure out the assignment of treatment easily. In other words, the purpose of using multiple sequences is blinding. Perhaps the best way to verify that ideal condition is by comparing a broad spectrum of baseline information collected at the beginning of every period. The following table illustrates a 2 x 4 crossover trial with baseline measurements:

	Period 1		**Period 2**		**Period 3**		**Period 4**	
Seq. 1	Baseline	A	Baseline	B	Baseline	C	Baseline	D
Seq. 2	Baseline	B	Baseline	C	Baseline	D	Baseline	A

Equivalence of baseline measures is reasonable evidence to believe the comparability of treatment groups between different periods.

Confounding is inherited in crossover setups. Of sequence, period, and treatment, the effects of any one confound with the interaction effects of the remaining two. Consider the 2 x 2 crossover setup illustrated in the following table:

	Period 1	Period 2
Sequence 1	X_1 A	X_3 B
Sequence 2	X_2 B	X_4 A

Let X denote the values at the combinations of sequence and period. A measure of treatment effects is

$$A - B = \underbrace{(X_1 + X_4) - (X_3 + X_2)}_{\text{Treatment}} = \underbrace{(X_1 - X_3) - (X_2 - X_4)}_{\text{Period-sequence interaction}}$$

which is also the difference of period effects between sequences, a measure of the effects of period-sequence interaction. Therefore, it is equally valid to interpret A – B as either the effects of treatment or the effects of period-sequence interaction. Similarly, the sequence effects confound with the effects of period-treatment interaction, and the period effects confound with the effects of sequence-treatment interaction, both of which can be simply demonstrated by rearranging the table. In the following table, for instance,

	Period 1	Period 2
Treatment A	**X_1 Sequence 1**	**X_3 Sequence 2**
Treatment B	**X_2 Sequence 2**	**X_4 Sequence 1**

it is easy to demonstrate that

$$\underbrace{(X_1 + X_4) - (X_3 + X_2)}_{\text{Sequence}} = \underbrace{(X_1 - X_3) - (X_2 - X_4)}_{\text{period-treatment interaction}}$$

If the patients assigned to treatment sequences are identical, which logically rules out any sequence effects, the period-treatment interaction effects may also be interpreted as carry-over effects, meaning that the treatment effects in one period still exist in the subsequent periods. Although zero drug concentration at the end of washout period is a strong evidence for small carry-over effects, it cannot be used as sole evidence to argue for small period-treatment interaction effects.

Therefore, crossover setups are useful only when convincing evidence is available from experiences or literature that the ideal conditions are generally met. When such evidence is not available or when substantial

patient withdrawals are expected after the first period, crossover setups should be avoided.

8.3 Randomization

Randomization is a technique to assign patients to treatment groups with an equal opportunity. The purpose is to make treatment groups comparable by evenly distributing the uncontrolled factors among groups.

Simple randomization is generally effective for this purpose when large number of patients are allocated to a small number of groups. By simple randomization, each patient is assigned an identification number, and this number is then randomly matched to the group identification numbers. When the number of patients in treatment groups is small, chances are that the groups may differ from each other considerably. Then randomization has to be balanced with respect to critical factors that may have substantial impact on the outcome.

By balanced randomization, newly recruited patients are classified by critical factors, and assignment is based on the distribution of existing patients in the trial. Suppose at a point of patient recruitment, 80% of the patients on active treatment have history of chronic ischemic heart disease while only 30% of the patients on placebo have the history. To balance the treatment and placebo groups in this regard, incoming patients with a positive history should be given a chance greater than 50% to the placebo group, and vice versa for patients with a negative history. The actual percentage for balancing depends upon how soon the recruitment quota is being met. Balanced randomization can only be dynamically carried out *during* the trial, and telecommunication is essential for fast access to a centrally controlled computer.

The outcome of randomization cannot be precisely predicted, and thus, the effectiveness of randomization should always be scrutinized. A good practice is to take baseline measurement before randomization and, after completion of patient enrollment, not necessarily completion of study, carefully compare the baseline measures among treatment groups. Nowadays, most randomization plans are generated with computer programs, taking into account study logistics such as drug supply, packaging, shipping, storage and dispense. The quality of computer

programs varies, so it is prudent to validate the program before patient enrollment.

8.4 Stratification

Stratification is grouping of patients by certain characteristics before randomization. The purpose is to make comparisons on similar patients. Stratification is a great technique to control the effects of critical factors on patients' responses so that the comparisons of primary interest can be made more precisely. However, it is also a restriction to randomization that may prolong the time of patient recruitment, increase study cost, and even jeopardize the study for an increased chance of unbalanced patient distribution within strata. The general principle is to use large strata without too much slowing down patient recruitment. The following table illustrates a parallel setup stratified by investigator center and the history of steroid use:

Center	Steroid Use	A	B	C
1	Yes	2	2	2
	No	2	2	2
2	Yes	2	2	2
	No	2	2	2

In each center, twelve patients need to be identified, and they are divided into two groups of six by the history of steroid use. The six patients in each group are then randomly assigned to treatments A, B and C, with two patients under each treatment.

Stratification slows down patient recruitment. A block of patients who meet the stratification criteria have to be identified before they can be randomized, and the block size has to be large enough to keep the evaluators blind of treatment assignment. Suppose the primary interest is comparing the effects of drug D and placebo, and the patients are stratified by the history of methotrexate use. The minimal block size is two patients; i.e., two patients on or not on methotrexate must consent to participate in the trial before randomization. However, because it is too easy to find out who is on drug D or placebo by randomizing just two patients, we may have to identify four patients and randomize them to drug D and placebo.

The problem that is potentially serious is that stratification may introduce confounding if the strata are small. The following table illustrates the problem:

	Drug A	Drug B
Center 1	**2 patients**	0 (withdrawn)
Center 2	0 (withdrawn)	**2 patients**

where four patients were assigned to Drugs A and B in each center. Later, each center had two patients withdrawn from the study, and they happened to be on the same drug. The consequence is that between the two pairs of available patients, A – B = C1 – C2; in other words, drug and center confound the effects of each other. Such unbalance due to patient withdrawals is unlikely to occur if there are a moderate number of patients in each center. However, if each center is required to recruit too many patients, some studies might take decades to complete.

Investigator center is the most common stratification factor in multicenter trials. The current practice is to initiate many centers simultaneously, with limited quota of patient recruitment to each center. The following table presents a typical patient distribution from study designed for comparing three treatments with 10 patients anticipated in each treatment group:

Center	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Drug A	x	x	x		x	x	x		x	x			x	x	x
Drug B	x	x		x	x	x		x	x	x		x		x	x
Placebo		x	x	x		x		x		x	x		x		x

In this table, x denotes available observations. Each center is required to identify three patients and given a block of supply of placebo and drugs A and B. The advantage with this setup is twofold. First, each center makes a relatively small contribution to the final result so that confounding from centers, if any, will be relatively small. Second, within-center comparisons can still be made for the majority of centers, so that the variation due to centers is more or less controlled. Problem with this setup is that the effects of center-treatment interaction may not be evaluable for all centers. The issue of evaluating center-treatment interaction is discussed in Chapter Five. A solution is to group centers by similarity in

medical practice. For instance, the fifteen centers may be grouped into five or six clusters, within each of which the medical practice is considered to be similar. Then the question is how to determine similarity. On the long list of factors that affect medical practice, the patient population and institutions where the physicians received his or her major training are perhaps most determinate.

8.5 Blinding

Blinding is a technique to keep the patients, investigators or both from knowing the assignment of treatment. The knowledge of treatment assignment can profoundly impact the patients' responses as well as the investigators' assessment of treatment effects. The purpose of blinding is to eliminate the potential confounding effects of that knowledge on the evaluation of treatment effects. Most highly regarded studies are double-blind studies, meaning that neither the patients nor investigators know the assignment of treatment until study completion and data freeze. After data freeze, no changes to the data are supposed to be made without careful documentation.

There are no fast-and-hard rules to determine the extent of blinding. It is largely determined by the appreciation of the effects of open labeling on the outcome, and for most of the part, it is a judgment call. From patients' perspective, the knowledge of therapeutic intervention may alter the patients' behavior, the perception of well being, and the seeking of alternative care. For patients, the adventure with an unknown therapeutic intervention creates a mixture of anxiety and hope, whichever is more depends upon the prognosis of the disease being treated. From investigators' perspective, conscious or subconscious discrimination of patients on different treatments at the stage of treatment administration and assessment of responses is perhaps the most significant confounding factor in interpretation of treatment effects. For investigators, experimentation with a therapeutic intervention is a mixture of curiosity and desire for success, and that driving force makes blinding an extremely tough job.

Extensive experience has accumulated for blinding in drug trials. In experienced laboratories, variety of compounds can be formulated into dispensable drugs of similar appearance, texture and taste that ordinary

people can hardly tell any difference. In spite of this, blinding may still be difficult to maintain for long if any of the comparing drugs can be identified by its unique clinical profile, such as outstanding efficacy or unusual adverse effects. Blinding is most difficult in trials involving procedures, such as massage, acupuncture, psychological counseling, group therapy and etc., and it is almost impossible in trials involving surgery, radiation, imaging and medical devices.

If blinding is difficult or impossible, a few other options are available to minimize the confounding from knowing treatment. One is to utilize what is generally believed objective parameters to measure therapeutic or adverse effects. Survival, for instance, enjoys wide acceptance as an outcome measure, not only because it directly translates into patient benefit, but also because there is not a whole lot of ambiguity when talking about death. Men can live a thousand lives, but never have a different death. Others, such as microbiology studies, biochemistry and cytology profiles, biopsies, hemodynamic measures and some image studies, generally have high reproducibility in similar clinical settings and through similar operating procedures. In acute settings, these parameters are less subject to the influence from patients' knowledge on therapeutic intervention, and the investigators have no direct control without giving concomitant treatments. Another option is to measure long-term, as opposed to short-term, outcome. Episodic issues and the intensity of intervention affect short-term outcome, whereas long-term outcome more depends upon the nature of the disease and the effectiveness of intervention in modifying the underlying disease process. Blinding is most effective in eliminating confounding due to the knowledge of treatment assignment for trials of therapeutic interventions that are equivocal or borderline efficacious and on indications that are waxing and waning in manifestation. Blinding is probably not necessary for trials of therapeutic interventions that are highly effective or on indications that are rapidly progressing with ominous prognoses.

Nothing can prevent fraud, however, unless the execution and evaluation of trial are completely dissociated from the consequence of trial result. Such dissociation is only achievable in large-scale trials with adequate financial support. Independent investigators may be employed and compensated for execution of study protocols. An independent,

service-oriented third party evaluation team is also valuable in monitoring ongoing trials.

In summary, blinding is important for the control of confounding due to the knowledge of treatment assignment and the consequent potential discrimination of patients on different treatments. However, blinding is difficult to implement. The standard practice that dissociates the beneficiary of trial result from the execution and evaluation of trial is generally very costly, and it is by and large an organizational behavior. For less funded studies where complete blinding is not feasible, the investigators need to perform careful interest analysis of participating parties, explore the impact of potential discrimination on study result, and to seek opinions from peers and regulatory agents for consensus.

8.6 Control

Control is the basis against which the effects of treatment are evaluated. There are two basic types of control: longitudinal control and parallel control. Parallel controls are categorized into placebo control and active control. Appropriate control is vital for the integrity of a clinical study program.

8.6.1 Longitudinal control

Longitudinal control is the measurement of responses at a time point, usually before the administration of treatment, and treatment effects are measured by comparisons to the control. If, for instance, the control is baseline measures before treatment, treatment effects may be simply measured by change from baseline. If, as another example, the control is responses at the early stage of treatment, the comparison of responses between the early and late stages of treatment measures the effects of treatment. With longitudinal control, treatment effects are measured by within-patient comparisons, meaning that the responses from the same patient separated over time are compared. The use of longitudinal control is profitable if the precision of within-patient comparisons is much higher than that of between-patient comparisons, the comparison of responses between different patients.

Time effects confound the effects of treatment as measured from longitudinal control. In clinical practice, time is always a critical factor in

making diagnosis and assessing the effects of therapy. Time cures. This is perhaps true for all self-limiting diseases. With time, the body utilizes its reserve to maintain life, builds up immunity and repairs structural damages. If treatment is given to patients with a self-limiting disease, the effects of treatment, measured by the difference before and after treatment, cannot be separated from the pure effect of time. Had the treatment not been given, the same difference might have been observed over the same period of time. Disease fluctuates over time. This is true for many chronic diseases that undergo remission and exacerbation over a long period of time, presumably due to the self-limiting nature of acute insults and adequate functional reserve of the body. Sometimes, the severity of illness during an acute insult is merely a manifestation of organ reserve. Any difference demonstrated with longitudinal comparisons within a relatively short period of time may just be a snap shot of the disease roller coaster. Therefore, the effects of treatment to induce remission, as measured by the difference before and after treatment, are confounded with the disease's natural tendency to remit after exacerbation.

Daily occurrence, like the following, demonstrates the importance of time effects. Anxious parents bring their child to clinic for wheezing, and the child is diagnosed for bronchiolitis probably secondary to respiratory syncytial viral infection and sent home with a nebulizer. The parents are told to bring the child back in two weeks. Ten days have passed, and the child is still wheezing. Now the parents become very skeptical to the first pediatrician. They then bring the child to another pediatrician in a highly regarded institute. The doctor does exactly the same thing except that he also prescribes a course of erythromycin. Two days later, the child stops wheezing and the parents choose the child a new pediatrician. Most likely, that erythromycin does nothing. The child stops wheezing because the disease is self-limiting, and twelve days are what it takes to resolve. While the second pediatrician might not offer a better management of the child, he is certain a master of the parents' psychology.

Treatment effects as measured from longitudinal control are also confounded with placebo effect. Placebo effect is a significant change from baseline observed from patients on placebo after being in trial for a period of time. The nature of placebo effect is rather complex. It may be due to behavior modification, such as better compliance with medications and cessation of cigarette smoking; it may be due to close surveillance and

timely treatment that are not otherwise available to general population; it may be due to the natural fluctuation of the disease being studied. Because patients might well improve for just being in the trial without active treatment, the claimed effects of treatment, as measured by any longitudinal, before-and-after-treatment comparisons, cannot actually be separated from that effect of placebo.

Because treatment effects based on longitudinal control confound with the effects of time and placebo, longitudinal control is most useful for studies in which the effects of time and placebo are minimal. Longitudinal control may be used in studies on steadily progressive diseases, diseases at their late stages when symptoms persist, and any other diseases with predictable course and prognosis. Chronic obstructive pulmonary disease (COPD) is an example. If the patients have persistent hypoxia, hypercapnia, and dyspnea, any consistent improvement in oxygenation and ventilation as measured by arterial blood gases and pulmonary function test are more likely attributed to the effects of treatment, not to the effects of time or placebo. This is because COPD at its late stage progressively deteriorates over time, and to date, nothing can actually reverse the progression. Congestive heart failure of all causes and primary pulmonary hypertension are other examples for which the effects of treatment may be measured against longitudinal control, because few things can consistently alter the course of these diseases.

8.6.2 Parallel control

Parallel control is an independent group of patients who are comparable to the group or groups of patients on treatment of primary interest. The idea is to let the patients in both control and treatment groups expose to otherwise the same condition over the time course of trial. The purpose of employing parallel control is to eliminate the confounding effects of time and placebo in the assessment of treatment effects.

Because parallel control is a different group of patients, the effects of treatment are evaluated by comparisons between control and treatment groups. As opposed to within-patient comparisons with longitudinal control, treatment effects are measured with between-patient comparisons. Comparability between control and treatment groups is vital for the logical validity of estimated treatment effects. Comparability is generally

achieved with effective randomization in the beginning of trial and proper blinding, if possible, during the trial.

The precision of measures in control group directly determines the precision of any comparisons with them. Therefore, it requires careful planning to set up parallel control. If the purpose is comparison, measures of control group ought to be as precise as measures of treatment groups. Researchers should resist the temptation of "savings" by reducing the number of patients in control group. If the quality of control is poor, the quality of any comparison with control goes down with it. In designing a series of studies, researchers need to exercise caution when planning small control groups for individual studies and hoping a large control group by pooling the study series . Few can argue against the fact that the same treatment may show different effects in different studies. Thus, the control group pooled from a series of studies may not be the same as a control group of the same size in a single study.

However, if the purpose is exposure, control groups smaller than treatment groups may be justified. Studies for safety generally require wide patient exposure to capture rare serious adverse events. For instance, if the incidence of a rare event is once every thousand, then on average, at least a thousand of patients are required for just a single catch. The size of control group may be determined by the minimal incidence of adverse events to be compared. If the purpose of control is to compare adverse events of minimal 5% incidence, for instance, then 200 patients in the control group may be sufficient for a reliable comparison of those adverse events. As long as the number of patients in the control group is sufficient for comparing adverse events of high incidence, the majority of patients should be assigned to treatment groups.

8.6.3 Placebo control versus active control

Placebo control sets a clean background. Just like white paper for colors and clear window for scenic views, placebo control allows for estimate of the pure and absolute effects of treatment. For a clinical study program, placebo control forms a basis for combining or cross-evaluating different studies. Problem with placebo control is that it may be difficult to design long-term studies if treatment drastically changes the course of disease in either direction. First, it would not take long for the evaluators

to figure out who is on what, and thus, it is going to be difficult to maintain blind for the study. Second, once a drastic difference is observed, it may not be ethical to continue placebo if treatment is beneficial, or to continue treatment even if it is temporarily harmful. For trial of angiotension-converting enzyme (ACE) inhibitor on diabetic nephropathy, for instance, if the trial is stopped because of the elevation of serum creatinine, the long term benefit in preserving kidney functions may not have a chance to show.

There are indications for which standard treatments are available, and it may not be ethically acceptable to conduct placebo-controlled studies. Then, active control, control with active treatment, is the natural choice. Compared to placebo control, active control sets a background at a relatively higher level. Just like color paper for colors and tinted window for scenic views, active control allows one to evaluate the relative effects of treatment. Active control can also form a basis for combining or cross-evaluating studies in a clinical study program, provided that the same active control is used throughout the program. Active control is favored in some clinical settings. First, when standard treatment is available, it would be most meaningful to compare any new treatment to that standard. Second, active control is more acceptable for patients, and therefore, it would be easier to get patients' consent for the study. A drawback with active control is that some well-known characteristics of standard treatment make it difficult to keep the investigators blind of treatment assignment.

The logical basis of utilizing control in evaluation of treatment effects remains the same no matter whether the control is active or placebo. The use of one as opposed to the other is entirely determined by the research purposes. In the current research practice, however, the use of active control is strongly discouraged. The rigid formality of statistical hypothesis testing in, based on the unrealistic statistical theory of Neyman and Pearson, can be blamed for this underplay of active controls. The absurdity of that theory is fully exposed in Chapter Ten. The statistical tests formulated under that theory could only claim significance for large differences with small p-values. Most of the time, comparisons to active treatment give rise to smaller differences than do comparisons to placebo. Because of the unfortunate adoption of $p \leq 0.05$ in decision-making, small differences no matter what and the resulting large p-values put researchers

in a disadvantageous position in making claims, even though evidence is otherwise strong for superior therapeutic effects.

Difference itself does not directly translate into anything. It is merely a formality depending on how the contrast is made. As we will clearly see in Chapter Ten, the statistical test of hypothesis is nothing but a crying game that these days researchers have to play for publications and making claims.

8.6.4 Pivotal control in clinical study program

A well-designed clinical study program should allow for integration of its studies to evaluate the consistency and heterogeneity of treatment. Appropriate control is vital for valid integration of clinical studies. The idea of pivotal control is using the same control to link all the studies.

The following table outlines three studies:

Study	Patient Population	Control	Treatment
1	Ischemic cardiomyopathy	Placebo	A, B
2	Ischemic cardiomyopathy	Placebo	A, B, C
3	Idiopathic cardiomyopathy	Placebo	B, C, D

Studies 1 and 2 are very much the same except that only study 2 has treatment C group. If the placebo groups are comparable with respect to all the measures, the two studies may be pooled together, and treatment group C may be directly compared to the combined treatment groups A and B. If the placebo groups are not comparable, which suggests some difference between studies, the two studies cannot be directly combined but compared. What are comparable is the within-study contrasts with placebo:

Study	Patient Population	Combinable contrasts
1	Ischemic cardiomyopathy	A – Placebo, B – Placebo
2	Ischemic cardiomyopathy	A – Placebo, B – Placebo, C – Placebo

The placebo control in these two studies is a pivot to connect the two studies. Because of the unique nature of placebo control, comparison

between placebo groups affords the most convincing evidence on the combinability of studies.

It is out of the question to directly combine study 3 with studies 1 and 2 because they are on different patients. However, it would be interesting to *compare* the effects of treatments in different patient populations. Treatment groups are not directly comparable across studies. Directly comparable are within-study contrasts with placebo:

Study	Patient Population	Comparable contrasts
1	Ischemic cardiomyopathy	A – Placebo, B – Placebo
2	Ischemic cardiomyopathy	A – Placebo, B – Placebo, C – Placebo
3	Idiopathic cardiomyopathy	B – Placebo, C – Placebo, D – Placebo

Within-study comparison to placebo adjusts both the effects of patient population and the effects of study implementation on the patients' responses to treatment. Once again, this placebo control sets a common ground for integrating the results of these three related but different studies.

8.7 Studies for dose-efficacy-safety relationship

Dose-efficacy-safety relationship is perhaps the most clinically relevant information for physicians to exercise clinical judgment in the care of individual patients. Information on efficacy and safety at a fixed dose in terms of averages is less useful for physicians to provide optimal care to individual patients. Deviation from recommended doses may render the physician liable to potential adverse consequences. A well established dose-efficacy-safety allows physicians to calculate the risk and benefit of treatment on an individual basis.

A dosing regimen must not be confused with dose, the mere quantity of the drug that the patient takes at a time. Clinicians concern more of dosing regimen than dose. A dosing regimen may be characterized with three parameters:

- drug concentration in serum or target organs,
- variation of drug concentration between dosing, e.g., peak and trough levels, and
- the time course of treatment, e.g., chemotherapy cycles.

Pharmacokinetic data and drug levels are necessary to define these parameters. Although drug levels are not necessarily clinically useful in predicting efficacy and safety on an individual basis, it is useful in clinical studies on groups of patients to define dosing, delineate what appear to be different regimens, and explain patient heterogeneity. It is impossible for a single business entity to exhaustively explore the efficacy and safety profiles of all possible dosing regimens and all possible combinations of those parameters at all possible levels. Once the industry provides the initial data, both the industry and medical society should carry the burden of continuous research on dose-efficacy-safety relationships. The purpose of this section is to discuss appropriate designs for studying dose-efficacy-safety relationships.

8.7.1 Confounding factors in evaluation of dosing

The biggest confounding factor is time. If a drug is taken over a period of time, the patients' responses, both toward and untoward, may be attributed to either the effects of dose or the cumulative effects of dose over time. If, for instance, a patient takes a drug at 10 mg twice a day over ten days, and then the same drug at 20 mg twice a day over another ten days, the responses at 20 mg may not be comparable to those at 10 mg. The argument is that 10 mg twice a day over 20 days may be just as effective, and hence, 20 mg, although well tolerated, does not necessarily produce more beneficial effects than 10 mg. In other words, the cumulative effects of 10 mg confound with the effects of a later dose increment. Therefore, unless there are substantial evidences that cumulative effects are negligible, studies with longitudinal control cannot separate the effects of dose changes over time from the effects of cumulative dosing.

Another important confounding factor is patient heterogeneity. Patients are all different, and so are their responses to any therapeutic intervention. This presents a problem to studies that adopt a dose titration schedule to a target. In these studies, the patients are started with a low dose, and then the dose is pushed higher and higher till the responses meet a pre-specified target criterion of efficacy or safety. The result of this dose titration is a spread of patients over the spectrum of their responsiveness to the drug. The titration tends to stop at low dose levels for patients who are sensitive to the drug for whatever reason. These patients meet the

response criterion before incurring further dose increment. On the other hand, for patients who are not as sensitive, the titration may go on to high dose levels till the patients meet the criterion or suffer intolerable side effects. Therefore, patients on high doses are not comparable to patients on low doses simply because their differences are a mixture of the effects of different drug levels and the heterogeneity in response to the drug.

Because of these confounding effects, studies for dose-efficacy-safety relationships should adopt a parallel design with the main focus on dosing regimens, not merely dose, a static quantity at a time. The purpose is to allow for a clear comparison of different treatment strategies that are implementable in clinical practice. A parallel design for dose escalating study is illustrated in the following table,

40 mg once a day	20 mg twice a day	10 mg twice a day

This study allows us to compare the three dosing regimens. A clear dose-response relationship is not only clinically useful, but also the best demonstration of efficacy. It is not necessary to always include a placebo control to demonstrate efficacy. Placebo means zero drug, which may be viewed just as another level of dosing. If no differences among the three groups are observed, a broader span of dosing may have to be explored to demonstrate efficacy, knowing that the range of dosing in any particular study may be completely off the optimal range where a dose-response relationship is present.

The most sensitive statistical analytic technique to present dose-efficacy-safety relationship is visualization of individual response profiles over the time course. Much of the information is lost when data analysts focus only on the mean responses at few static time points, and statistical tests have absolutely no role. Most clinically useful insights are gained through thorough investigation for cause of heterogeneity. An effective analytic strategy for this purpose is to group the patients by their responses, and then characterize the groups with respect to both pharmacokinetic measures and clinical features.

8.7.2 Dose escalating studies

Dose escalating studies are often phase II trials to find a safe and effective dosing regimen. The usual practice is to push the dose, if safety warrants, for the maximal therapeutic effect in a series of studies. However, trial initiation is a time-consuming and expensive process. Externally, protocols must circulate around through regulatory agencies if necessary, outside experts, site investigators, and Institutional Review Boards (IRBs); investigators and their staff must be selected, visited and trained. Internally, guidelines and procedures in all aspects of the trial must be initiated, documented, reviewed, validated if deemed necessary, and finally approved; collaborations must be established and tested among clinical department, data management, study logistic, clinical laboratories and investigator sites. There are circumstances, however, when programming a single dose escalating trial can save tremendous time and manpower in trial initiations. Effective telecommunication techniques make this feasible.

The following is a heuristic design, not to be copied in practice. The purpose is to illustrate how a slightly complex design may be programmed in a flexible organization with the aid of telecommunication techniques.

Suppose we have known that the drug at dose 5 mg is safe but has no clinically useful therapeutic effects. The next step would be to increase the dose and monitor the safety. If the increased dose turns out to be safe, the dose may be further increased till a plateau in therapeutic effects is demonstrated. This idea is illustrated in the following table:

Ratio:	1	2	2	2	2
Initial dose	**Placebo**	**10 mg**	**20 mg**		
First increment	**Placebo**		**20 mg**	**30 mg**	
Second increment	**Placebo**			**30 mg**	**60 mg**
Number of patients	3 x 10	2 x 10	4 x 10	4 x 10	2 x 10

The initial assignment of patients follows 1:2:2 ratios to placebo and drugs at 10 and 20 mg doses. If no serious adverse events are detected, newly recruited patients are assigned to placebo and higher dose groups in the same ratios. There is a placebo group at each step of dose escalation. This setup guarantees the comparability of the groups at each step of dose escalation. The highest dose group at each dose

increment is always repeated at the next step of dose increment. The purpose is to accumulate more observations for reliable estimates and to increase the chance to pick up adverse events of low incidence. Of course, if there is no interest in low dose groups, it is not necessary to repeat them. Dose advance may be conservative at the beginning and a bit aggressive if a safety profile starts to emerge. Fixed ratios for patient assignment to treatment groups are preferred because they simplify management and reduce the chance for errors. However, because we often do not know in advance exactly when the escalation stops, it would be prudent to assign more patients to placebo initially. The purpose is to guarantee an adequate number of patients in the placebo group when escalation stops unexpectedly.

The most critical decision is when the dose can be safely increased. This has to do with the number of patients exposed to the drug. After the first dose increment, as shown in the table, 60 patients have exposed to the drug. The study at that point has the sensitivity to detect adverse events of minimal incidence of 1.7% (1/60). For dose-dependent adverse effects, the study at the highest dose of 20 mg has the sensitivity to detect adverse events of minimal incidence of 2.5% (1/40). An independent safety review committee is necessary to constantly monitor the emerge of adverse events and to judge whether or not patient exposure is sufficient to comfortably advance the dose.

A dose escalating study like this greatly increases the difficulty of study logistic. The drug at a series of doses has to be formulated, and the quantity has to be large for quick shipping upon requests from investigator sites. If the drug is very expensive, the cost of the drug may not justify such a design. This design requires an independent safety review committee who have constantly access to the data and flag at the programmed turning points for dose increment. This design also requires dynamic randomization that fast telecommunication is essential for investigators to access the randomization code and receive the correct shipment of drug supplies. Despite these managerial complexities, the study is indeed programmable within a flexible organization. All activities are limited to the matching of randomization code to the correct drug supply. The investigators see no difference from a regular parallel study except that they need to make more phone calls to get randomization code and shipment of the drug that matches the code.

8.7.3 Reflection on the traditional phases II and III paradigm

The clinical development of an investigational drug (IND) is divided into phases II and III. The focus of phase II studies is exploring the safety and efficacy profiles, and particularly, identifying a tolerable dose that may achieve the maximal therapeutic effects. The results of phase II trials are consolidated, and a go or no-go decision is made to cast the die for a large phase III trial, the result of which, if favorable, is used to support an new drug application (NDA).

Question is whether a large phase III trial generates more clinically useful information than a series of quality phase II trials of high quality. In my opinion, most of the current phase III trials are merely designed to fit in the rigid formality for statistical test of hypothesis and to obtain small p-values. Usually, the information collected in phase III trials is not much different from that in phase II trials. The design is, however, geared for obtaining small p-values. To obtain small p-values with the usual statistical tests, the difference of the means between comparing groups has to be large, and the standard error has to be small. It is not uncommon to see that in phase III trials, the largest tolerable dose is compared to the least efficacious control so that the difference can be large. The sample size is fairly large in phase III trials. When the standard deviation is stable with certain number of patients, further increment of sample size reduces the standard error.

What information is clinically useful? Clinically useful information should help calculate the benefits and risks of a therapeutic decision on individual patients. Information on dose-efficacy-safety relationship and characterization of responders and nonresponders are helpful. What study result is most convincing? In my opinion, result consistently demonstrated in a series of studies is much more convincing than result flashed in a large single study. I recommend drugs that come in with clinically relevant information consistently demonstrated in a series of studies with logical designs and quality data. I am skeptical of drugs that merely associate with statistically significant averages in a study of whatever a size.

8.8 Studies for treatment substitution

Study for treatment substitution is to see whether or not a new therapeutic intervention can partially or completely substitute for the existing one. An interesting aspect of this design illustrates the concept of sensitivity and confounding from patient heterogeneity. The study is usually conducted when standard treatment is available, and the new treatment is either an adjunct to or a potential substitute for the standard treatment. A typical clinical scenario is the management of allergic and autoimmune disorders where cytotoxic drugs, like azathioprine, cyclophosphamide, methotrexate and cyclosporine, are constantly being attempted to reduce or spare the use of corticosteroids. Those studies are often referred to as steroid-sparing studies. Studies for treatment substitution provide not only evidence on the efficacy of new treatment but also valuable guidance on treatment transition.

8.8.1 Target range of disease control and study sensitivity

Typically, patients selected for study are already on standard treatment, and a target range of disease control is specified. The patients are randomly assigned to a control and the new treatment for substitution, and they are followed regularly over a period of time. At follow-up visits, disease control is assessed, and the underlying standard treatment is adjusted to maintain disease control within the specified range. The usual practice is to wean the standard treatment if disease control is within the specified range, and to intensify the standard treatment quickly if the disease flares or exacerbates beyond the specified range. The end points are the magnitude, timing and consequences of treatment substitution.

It is interesting that the sensitivity of the study, to some extent, depends upon the specified range of disease control. If the range is wide, the disease is allowed to fluctuate within that range without triggering the adjustment of standard treatment. Therefore, as measured by change of standard treatment, the study will not be sensitive to detect the effects of new treatment if they are weak or borderline. The weak effects of new treatment may be demonstrated by a better disease control. However, any change in disease control beyond that specified range will trigger the adjustment of standard treatment. Thus, the

difference in disease control cannot be greater than the limits set by that specified range.

Suppose cyclosporine is being studied for its steroid-sparing effect, and the criteria for steroid reduction are clearance of proteinuria and clear chest x-rays. Unless cyclosporine has dramatic effect to clear up inflammation, this study may not be able to demonstrate any steroid-sparing effect. Although the negative result might suggest a complete lack of efficacy with cyclosporine, it is well possible that the criteria are not sensitive enough to pick up its real beneficial effects. The benefit may be found by comparing disease control. We may find that more patients on steroid and cyclosporine combined therapy have less protein in their urine and less severe pneumonitis as shown by high resolution chest CT than patients on steroid only.

The above example shows how insensitive criteria in assessing disease control allow for wide changes without triggering steroid adjustment. Studies for treatment substitution are most sensitive when the range of disease control is tight so that any small change in disease activity duly translates into adjustment of the underlying treatment to be substituted. Problem is that if the range is too tight to allow for clinically acceptable disease fluctuation, the study may end up with clinically irrelevant result. Suppose now, for the same steroid sparing study on cyclosporine, the criteria for tapering steroid are improvement of constitutional symptoms, quantified reduction of immune complex deposition and reduced gallium avidity in lung scans. With these criteria, the effects of cyclosporine may be shown with respect to both reduction in overall steroid use and a better control of the disease.

8.8.2 Patient heterogeneity and disease fluctuation

Because patients respond to treatment differently, dose adjustment by achieving a target disease control spreads out the patients over a spectrum of responsiveness. Dose reduction tends to occur in patients who readily respond to the treatment, whereas dose increment tends to occur in patients who are not sensitive or responsive. Therefore, it is important in study for treatment substitution that the patients are divided into responders and nonresponders. Characterization of responders and nonresponders helps to gain insight to the clinical problem.

Another fact that design of studies for treatment substitution needs to take into account is that disease fluctuates over time. Remission prompts a better disease control and results in dose reduction, whereas flare or exacerbation prompts a worsening disease control and results in dose increment. These adjustments cannot be attributed to the effects of treatment, but the natural history of the disease. This confounding effect of disease fluctuation precludes the use of longitudinal control in such studies. Parallel control is necessary to demonstrate the true therapeutic advantage on top of disease fluctuation.

8.9 Determination of sample size

Adequate quantity of observations is essential for reliable conclusion. The more we observe, the more we learn. There is no gold standard to judge how many observations are adequate. Resources generally set the top limit. Only occasionally, a lower limit may be set for reasons.

That the sample size of a study can be determined with statistical power is entirely a utopia built on the unrealistic statistical theory of Neyman and Pearson. Power is defined as the probability of rejecting a hypothesis correctly. Although sounds attractive, power is not implementable with observable measures. The question is that no one knows if we have done the right thing without testing it in future studies. True power requires the knowledge of truth to make judgment. In experimental sciences, truth is not readily available, but what to be found out. Because no observable information is available to compute power, Neyman and Pearson utilized mathematical distributions as a surrogate of truth to judge the decision on rejection or acceptance of a hypothesis. Because statistical power is not measurable, the sample size determined with power has absolutely no touch to reality.

There has not been much development on sample size estimation on a realistic basis. This section will explore few ideas and develop the corresponding measures for a rough estimation of sample size. The criteria used are sensitivity, stability of measures, target standard errors, and information. The idea of information is due to Ronald A. Fisher.

8.9.1 Sampling unit

Sampling unit is the accounting unit for the quantity of observations and is entirely determined by the research purposes. In general, the sampling unit for a study should be the accounting unit of the population to which the result of study is intended to apply. For case studies where the primary interest is specific patients, the sampling unit should be one piece of information. Therefore, resource should be spent on exhaustive search for information on those specific patients. For most clinical studies where the primary interest is patient population, the sampling unit should be one patient, and ideally, that sample should be representative of the patient population. Therefore, spending of resource should be balanced between adequate number of patients and sufficient information from each patient.

The question constantly facing all clinical researchers is whether to spend limited research funding on exhaustive study of few patients or a focused study of a good number of patients. Just as a better way to see New York City is walking on the street, not scrutinizing every brick of a building, a better way to understand a patient population is to see as many patients as possible, not every cell of one patient. A common mistake is generating a good deal of data from exhaustive examination of specimens from few patients while the research purpose is to get a result for the patient population. That detailed knowledge about those few patients may cost a great deal, but is of little use to draw conclusions on that patient population.

8.9.2 Criteria for projecting sample size

Sensitivity is a criterion for setting the minimum number of observations. A measure of sensitivity is the reciprocal of the number of observations:

$$\text{sensitivity} = \frac{1}{\text{number of observations}} \times 100\% .$$

If the incidence of an adverse event is 1%, for instance, the sample size, on average, needs to be at least 100 in order to expect a catch of that event. A

study of 100 patients, hence, has the minimal sensitivity for detecting adverse events whose incidences are 1% or higher.

If the response measure is ratio or percentage, the sensitivity measure also indicates the robustness of that measure against the change of a single observation. Suppose 50% is the percentage estimated from 10 observations, i.e., 5/10 = 50%. If the value of one observation is different for any reason, the result would be either 4/10 = 40% or 6/10 = 60%. As such, the percentage could change by 1/10 = 10% upon the fate of a single patient. Had the percentage been estimated from 100 observations, that percentage would have wobbled only from 49% to 51% upon the change of a single observation.

Stability is another criterion for setting the minimum number of observations. If a series of studies on the same topic are available, one may select critical measures for the response parameter of interest and compare their magnitudes among studies of different sample sizes. In the following graph, five summary measures are presented along studies of increasing sample sizes:

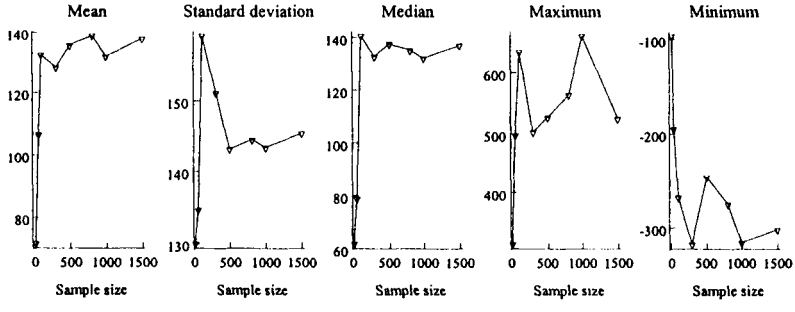


Figure 8.1 Fluctuation of summary measures by sample size

When the sample size reaches between 100 to 200, the mean and median become fairly stabilized. The standard deviation is not stabilized, however, until the sample size reaches 500. Further increase of sample size does not significantly change these measures. Therefore, the sample size may be chosen to be 150 patients. With this number, we may expect a

mean or median that fairly represents the patient population without compromise of precision.

If a series of studies are not available, a single study on the study topic may be useful. To study stability, we first identify the summary measure of primary interest and then watch the changes of that measure upon sequential additions of a single observation. The projected sample size is where the changes converge to zero. The following graph presents the changes of mean from a 750-patient study:

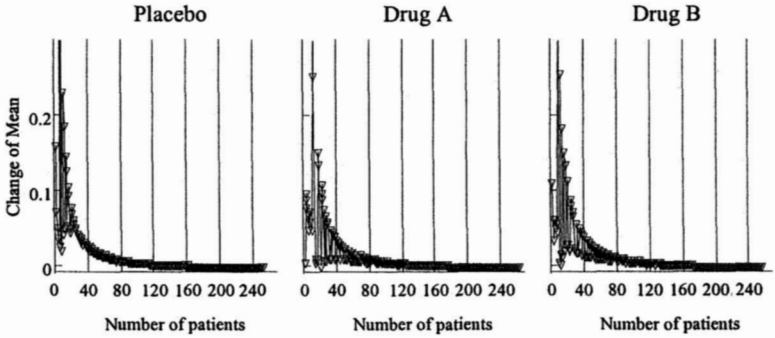


Figure 8.2 Changes of mean with increasing sample size by treatment

The observations are sorted in descending order by their absolute deviations from their respective group means. The sequential changes of mean in each group upon the addition of one patient with a smaller deviation at a time are computed by

$$\text{change of mean} = \text{mean of first n observations} - \text{mean of first n} - 1 \text{ observations},$$

where the n^{th} observation has a smaller deviation from the group mean than the $(n-1)^{th}$ observation. The graph shows that when the sample size reaches 40 – 80 for each group, the means of all three groups become stabilized. Therefore, the sample size of 60 patients in each treatment

group appears to be adequate if we use the means to characterize the patient population.

The sample size estimated from a single study with this method should be used with caution. First, the estimate is rather conservative. The large fluctuations in the beginning are artificial. Observations are sorted by their deviations from the group means, which is meant to create the worst scenario. For real data, such an order does not exist. Second, the estimate heavily depends upon the exposure and sample size of source study, and therefore, it could be misleading if the source study does not represent the patient population.

The drawback from arbitrary sorting of data may be totally avoided if we can use re-sampling technique. The idea is to draw a series of samples of increasing size, with or without replacement, from patients in the source study, and then compare the distributions of these samples. The number beyond which further increment of sample size does not significantly alter the distribution may be chosen to be the estimated sample size.

Careful review of patients in placebo groups in completed studies may shed light in planning studies in similar patient population. This is particular useful when no information *a priori* is available for the treatment to be studied. Suppose five studies have placebo control on similar patient populations. If a consistent profile is demonstrated in three out of the five control groups, the sample sizes of those three control groups provide a valuable reference to determine the sample size for future studies on the same patient population. Suppose the sample sizes of those three control groups are 90, 150 and 200. Then a sample size of 100 patients may be set as the minimum for a study in this patient population. The idea is this: Although the actual responses to treatment may be diverse and more patients are required to characterize them, at least 100 patients are necessary to just present a picture of that patient population that is robust enough to be consistently demonstrated in a series of studies.

Finally, the precision of mean, as measured by its standard error, can be a criterion. Suppose that a previous study has sufficient number of observations, so that the mean and its standard deviation for a critical measure are rather stabilized. The sample size for achieving a desired standard error for the same measure in a similar study may be projected by

$$\text{sample size} = \left(\frac{\text{standard deviation from the previous study}}{\text{desired standard error for the planned study}} \right)^2 .$$

This utilizes the functional relationship between the standard error and standard deviation:

$$\text{standard error} = \frac{\text{standard deviation}}{\sqrt{\text{number of observations}}} .$$

If the standard deviation from a large study is 100, and the desired standard error is 5, for instance, the projected sample size would be $400 = (100/5)^2$. The same calculation may also be carried out in terms of information, which is the reciprocal of squared standard error:

$$\text{information} = \frac{\text{number of observations}}{(\text{standard deviation})^2} = \frac{1}{(\text{std})^2} \text{ x number of observations.}$$

$(1/\text{std})^2$ is the unit information contributed from every single observation. From the previous study, the unit information from each patient is $0.0001 = (1/100)^2$. To achieve information of $0.04 = (1/5)^2$, the number of patients should be $400 = 0.04/0.0001$.

Because the standard error is used to quantify the quality of summarization in many analytical procedures, one may set forth the most desirable result from an analytic procedure and work backward to find out what the standard error ought to be. Once the desired standard error is defined, it is straightforward to compute the sample size. For instance, a small p-value is expected from the analysis of variance if the difference of means is twice as much as the standard error. If the difference is 10, the standard error should be at most 5 in order for the ratio to be greater than 2.

The problem with the standard error criterion is that the standard error itself is a questionable measure with no straightforward interpretation. The practical meaning of small standard error is rather obscure, and claims

based on small standard errors can be seriously misleading. Although it may be used to evaluate the efficiency of clinical studies, the standard error should not be used solely to determine the sufficiency of patient exposure.

8.9.3 Measurement of efficiency

Efficiency is the dollars spent for unit information. While dollars are tangible and their face value is easy to measure, information is not. Here, information is defined as

$$\text{information} = \frac{\text{number of observations}}{(\text{standard deviation})^2} = \frac{1}{(\text{standard error})^2}.$$

This narrow definition of information does not seem to be immediately helpful for the design of clinical study. However, this measure can be useful for evaluating the designs of completed studies. The experience from completed studies may be invaluable for the success of future studies.

Suppose each patient costs \$3000 in a simple randomized study and \$3200 if the patients are stratified by their prior treatment. Is the money well spent on stratification? To answer this question, we may choose few critical measures from which the conclusions are drawn, and then perform the analyses of variance for treatment effects with and without the stratification factor. Suppose 100 patients were enrolled in the study, and the standard errors for a critical measure are 5 with stratification, and 7 without stratification. Now we compare the dollars spent on unit information:

$$\begin{aligned} \text{Without stratification:} \quad & \$3000 \times 100 / (1/7^2) = \$14{,}700{,}000; \\ \text{With stratification:} \quad & \$3200 \times 100 / (1/5^2) = \$8{,}000{,}000. \end{aligned}$$

Although the spending is \$200 more on each patient with stratification, the gain of information was so great that the spending on unit information is only an half of that without stratification.

Researchers must not be horrified by those astronomical numbers after the dollar sign. Those numbers are only meaningful for comparison

purposes. Unit information appears to be rather large a unit. Microunit information may be more convenient, where microunit information is 1/1,000,000 of unit information. Now if we repeat the previous calculation in terms of microunit information, we should feel much more comfortable with the following less spectacular numbers on the earth:

$$\text{Without stratification: } \$3000 \times 100 / (1{,}000{,}000 / 7^2) = \$14.7$$
$$\text{With stratification: } \$3200 \times 100 / (1{,}000{,}000 / 5^2) = \$8$$

The calculation speaks out that the cost of microunit information is $8 with stratification, and $14.7 without. Therefore, stratification for this study is money well spent.

9
Integration of Clinical Studies

Summary

This chapter focuses on the principles and techniques for integration of clinical studies. The whole idea is assessing the quality of clinical studies and putting together related clinical studies for a comprehensive account of a therapeutic intervention. A set of parameters are parceled out for assessing study quality, and they are sample size, comparability among treatment groups, time course, precision of estimates, and documentation. The behavior of medical publication is analyzed to gain insight to the information contained in medical literature. The techniques of data analysis are designed to demonstrate the consistency and heterogeneity of study results against measures of their quality, and the main techniques are data visualization and the analysis of variance. A general analysis of variance technique is developed for an integrated analysis of multiple studies with the maximum likelihood technique of Ronald A. Fisher. The last section is devoted to the current statistical methods on meta-analysis.

9.1 An overview

The efficacy and safety of a therapeutic intervention generally need to be evaluated in the following aspects:

- the consistency across different patient populations and care givers,
- the relationship between the intensity of therapeutic intervention and the magnitude of response, i.e., the dose-response relationship, and
- the long-term impact on the patients being treated.

Due to the complexity of human subjects and limitation of resources, it is impossible to design a single study to address all these aspects at the same time. The current practice is to design simple studies each to address a specific question with data of high quality. Therefore, a series of clinical studies are necessary for a comprehensive evaluation of every therapeutic intervention.

A comprehensive evaluation of therapeutic intervention may be accomplished in two different approaches. The ideal approach is going through a rigorous clinical study program that consists of a series of carefully planned studies. The integrity of a clinical study program requires a focused target, strategic planning, and quality control. The success of a clinical study program depends upon medical insight, adequate funding and managerial skills in a regulated environment. The clinical development programs in the pharmaceutical industry are mostly of this type. They are not different in any way from any other scrupulous commercial business.

The other approach is by review of literature. To gain insight into medical literature, careful study of the behavior of medical publication is required. Section 9.2.3 is devoted to this matter. By and large, contributions to medical literature are made by individuals, often physicians, who are interested in research, review the literature, identify a problem yet to be satisfactorily resolved, propose a study to address the problem, and, if funded, carry out the study. Compared to studies in a well designed clinical study program, the focus is broad, the quality varies, and the conclusion heavily depends on the reviewer, who may be biased, consciously or subconsciously, in selecting studies from the medical literature.

In recent decades, the volume of medical literature has exploded. In line with this is call for a systematic approach to organize and evaluate this ever expanding body of information. The development of meta-analysis is a response to this call. A widely quoted definition of meta-analysis is a statistical technique for *quantitative* integration of clinical studies.

Although it is not lack of serious doubt whether or not those statistical concepts and techniques in meta-analysis are truly useful, the idea of integration is indeed inspiring, and the pioneers of meta-analysis deserve credit.

This chapter focuses on some of the key issues on evaluation and putting together of clinical studies. I benefit a great deal from some of the interesting discussions put forward by several panels of experts in meta-analysis. The evaluation of clinical studies and the analysis of publication behavior are very much in line with those already discussed in meta-analysis. The technical development, however, takes a different approach for a completely different interest. Instead of weighted averages over a hodgepodge of studies, the primary interest is to demonstrate the consistency and heterogeneity of study results against measures of their quality. The techniques are, therefore, data visualization and the analysis of variance with respect to measures of study quality. When the original data are available, a generalized analysis of variance technique is proposed for an integrated analysis of multiple studies with the maximum likelihood technique of Ronald A. Fisher.

9.2 The principle of integration

Integration of clinical studies requires meticulous planning, and it presents an intellectual challenge. The points discussed in this section may help researchers to form a framework in putting together studies. Those points are gathered with respect to study objective, parameter selection, and source of information.

9.2.1 Consistency and heterogeneity

The sole purpose of putting together clinical studies is to demonstrate both the consistency and heterogeneity of patients' responses to a therapeutic intervention. Consistency is repeated occurrence of the same event in response to an intervention in the same setting. In expedition to the unknown world where no experience *a priori* is available to judge whatever occurs to us, consistency is perhaps the single most important criterion to judge the meaning of our findings in that world and to guide our operation to that world. Indeed, it has been a respectful practice in

research society to repeat studies to verify any critical result before further actions take place. With no exception, consistency is one of the necessary criteria to establish the effects of a therapeutic intervention, and studies showing a consistent result afford the strongest evidence for making claims.

On the other hand, patients are all different. Two patients never respond to the same treatment exactly the same. Heterogeneity is, therefore, inherited in all clinical studies, and it represents the rich information brought from studying groups of patients. In fact, studying heterogeneity is not only challenging but also most rewarding in clinical research. Only by studying heterogeneity are we able to identify important factors that exerted significant effects on the patients but had never been recognized prior to the study; only by studying heterogeneity can we fully appreciate the effects of therapeutic intervention. Immediately identifiable causes of heterogeneity include study design, patient population, care giver, treatment, and the time frame of data acquisition. Most of the heterogeneity cannot be explained, however, and is ascribed to the uncontrolled causes.

In the statistical literature on meta-analysis, there are authors who advocate that the purpose of meta-analysis, or integration of clinical studies, is to increase "statistical power" in order to gain a "statistically significant" result from studies whose results are not "statistically significant". To understand what this advocation really means, we have to know the analytic techniques proposed by those authors. Almost all the statistical testing procedures in the current statistical literature on meta-analysis are based on the means. Most of the time, these means are gathered from published articles, they are averaged over selected studies, and the averaging is weighted by some arbitrary parameters that are believed to represent the contributions from those selected studies. Generally speaking, the statistical test on this weighted mean is to compare it to its standard error. The calculation for standard error is discussed in the last section. In summary, the primary interest of statistical testing in meta-analysis is a weighted mean and its standard error. Comparing this mean to its standard error yields a p-value, considered to be "statistically significant" if it is small or "statistically insignificant" if it is large.

In my opinion, this advocation of statistical testing in meta-analysis is leading researchers astray from the genuine interest of clinical research, and its absurdity lies in the unrealistic statistical theory of Neyman and Pearson and the entire fabrications of statistical testing flowing out of that theory. Interested readers may go to Chapter Ten for a more detailed account of that theory and statistical testing. My point is that statistical testing by comparing means to their standard errors does not help us understand the data and clinical problems in clinical studies. Never can a mean fully represent the rich information in a clinical study, and rarely are clinical studies exact replicates. A mean over a hodgepodge of heterogeneous studies makes no practical sense. Such a mean may be used, at most, as a reference point to measure heterogeneity. There are times when pooling of studies may give rise to a more reliable estimate of the magnitude of therapeutic effect. Those times are when the studies to be pooled demonstrate a consistent therapeutic effect. Even then, the main focus should still be the consistency and heterogeneity. It is because the scientific objective of clinical research is to understand the clinical problem. Studying consistency and heterogeneity helps us to achieve that objective, while a narrow focus on the means takes us to nowhere.

9.2.2 Parameters to consider in planning integration studies

Different integration studies have different plans for achieving different study objectives. However, the following considerations are probably essential to every integration study.

The first is the nature of therapeutic intervention to be evaluated. For the purpose of integration, a therapeutic intervention can be defined either as a specific treatment, such as a single drug at a specific dose, or as different treatments for the same therapeutic objective, such as modification of cardiovascular risk factors by all means. The goal is to identify the common ground on which the studies can be sensibly integrated.

The second consideration is the patient population. A patient population can be defined with a variety of parameters. Demographic parameters are certainly important. Past medical history, established prognostic factors, performance status, severity and chronology of illness and perhaps socioeconomic condition may also be pertinent parameters for

stratifying patients. The goal is to determine a set of inclusion criteria for the study and to stratify the included patients for a better precision during data analysis.

The third consideration is the quality of clinical studies. Garbage in, garbage out. A single study of poor quality may jeopardize the entire integration study by introducing confounding and destroying the precision of estimates. The rule of thumb is that studies of different qualities should not mix. The quality of a clinical study may be judged by two different standards. One is novelty, and the other is reliability. Studies that present novel ideas, reveal problems, share experiences, and provoke research initiatives are extremely valuable. However, these conceptually interesting studies do not necessarily fit in an integration study where the main objective is to seek consistency and investigate heterogeneity. The reason is fairly intuitive that while solid blocks build the Great Wall, fictions can only make up a ghost. An integration study is most useful when each of the included studies affords reliable information. Section 9.2.4 discusses some of the criteria for selecting reliable studies. The point is to keep an open mind to great ideas, but pick only reliable studies for integration.

The final consideration is the choice of endpoint parameters. For integration of clinical studies, it is vital to define the endpoint parameters without ambiguity. The goal is to make sure that no oranges are mixed in when the interest is apples. While some endpoint parameters are well defined, for instance, 5-year survival rate, many require elaboration. Belonging to that many are quality of life, response rate, symptoms and composite measures. When dealing with these parameters, one should meticulously study every single element of a composite parameter, the timeframe involved, the questionnaires used, the grading of symptoms, and both the numerator and denominator of a reported ratio. Sometimes, a phone call away is just what it takes to clarify subtle issues that cannot be fully appreciated by merely reading the article. Sometimes, the parameters are so intrinsically complex, as ejection fraction, intracranial pressure, or measures of a dynamic process, that they have to be appreciated in the specific pathophysiological setting. Socioeconomic parameters are even trickier. Hospitalization, for instance, is entirely depending on the admitting procedure of the hospital. Studies that utilize

these "soft" parameters generally show great heterogeneity, and the conclusions are often open to dispute.

Modern research protocols often set an arbitrary hierarchy for endpoint parameters. The hierarchy is usually a pyramid, consisted of a single primary endpoint parameter on the top, followed by groups of secondary and even tertiary endpoint parameters. The rationale behind such a pyramid is to avoid multiplicity, an issue often raised by statisticians based on the unrealistic statistical theory of Neyman and Pearson. The absurdity of that theory is fully exposed in Chapter Ten. What I want to point out here is that the entire hierarchical structure of endpoints is a hallucination, and multiplicity is completely a non-issue. The practical meanings of endpoint parameters and their quality are most important. Not all endpoint parameters are interesting, and not all endpoint parameters translate directly into patient benefit. Researchers should have the flexibility to choose whatever parameters of interest, and must never be distracted from genuine scientific interest by any pyramid based on merely a statistical ground.

9.2.3 Source of information

Except when an appropriately designed clinical study program is available, the main source of information for integration studies is the medical literature. An analysis of the authors' interest and a review of the convention in medical publication would help us gain insight into what we get from that source. The main objective of this section is to analyze of the behavior of current medical publication.

Publication promotes recognition. The quantity and quality of publications are an important, sometimes the sole, measurement of achievement. "Publish or perish" best characterizes this tie of authors' personal interest to publication. The consequence is that only publishable studies are conducted. The shortage of long-term studies for chronic diseases is an example. That job takes a young physician, who has to live and hold the position long enough, and is willing to do tremendous work, not immediately visible, for a hopeful publication before retirement. This lack of incentive from publication perhaps explains the shortage. The only solution to this problem that I can see is making the job a commercial business for organizations. The point is to compensate for clinical

research of great public interest, but not immediately visible by publication, with direct financial benefit.

Publication presents happy endings. It is a well-recognized phenomenon that studies with positive results are more likely to be published than studies with negative results. People like happy endings, and medical publication is not different from Hollywood in this regard. This selective publication of positive studies introduces bias and creates, what is referred to in meta-analysis literature, the iceberg phenomenon. For scientific purposes, studies of high quality are equally important regardless of their results. Even studies without a result, the so-called failed studies, can offer valuable experience that increases the chance of success in future studies. In recognition of this publication bias, the importance of collecting both published and unpublished studies cannot be overemphasized for an objective and balanced review of available evidences.

Statistics is misused to determine happiness. While the Hollywood goes by basic instinct to determine happiness, medical publication uses the p-value of a statistical test. It has been a disaster that $p \leq 0.05$ is adopted to determine significance in medical publication. Most statistical analyses used in medical literature fall in the broad category of analysis of variance. With the analysis of variance, $p \leq 0.05$ merely means that the average is about twice as much as its standard error, no more and no less. Does $p \leq 0.05$ suggest a degree of certainty to reject or accept a hypothesis? No. Does it imply the logical validity of study result in terms of control of confounding? No. Does it convey in any sense the quality of study? No. Does it even a good measure of the precision of mean? The answer is, again, no. Because statistical testing is such garbage and has no practical meaning whatsoever, statistical significance must never be used as a criterion to select and judge clinical studies.

Quality control is not required for publication. For integration of clinical studies, we are most concerned with the quality of published studies. The key in quality control is access to, and independent audit of, the data. While data access has been an extremely sensitive issue and is difficult to obtain, auditing can be very expensive. This is probably why most of medical publications rely solely on the faith to authors than a valid quality control process. The consequence is that the published results may

be fraudulent, inaccurate, fragmented, biased and unreliable, and queries to the authors can be time-consuming, unwelcome, and expensive with no guaranteed payoff.

Novelty and quality, in my opinion, should be the sole criteria for medical publication. Novelty provokes controversy, and quality safeguards decision-making. Both are healthy signs of scientific research. Publication should not be just story telling, but an invitation to the data generated from the study. The agreement for scientific publication should include terms for free access to the data. The idea of database network or web under a global standard of clinical data management is appealing. A global standard will greatly facilitate data sharing and quality control. Discussion on this matter is continued in Chapter Twelve.

9.3 The quality of clinical studies

Appraisal of clinical studies is difficult to make with clear-cut guidelines, simply because it cannot be dissociated from the problems under study. Each specialty has to develop its own criteria to determine the quality of their studies. Here, we merely discuss the principles of quality assessment for this generic clinical study. The patients are screened for eligibility. After a run-in period, if necessary, eligible patients are assigned to treatments and followed, usually at a set schedule, over a period of time. The patients' responses are assessed in the period of follow-up. A study like this is assessed from three perspectives: the integrity of study design, the quantity of observations, and the completeness of study documentation. The central concept is the control of confounding.

9.3.1 The integrity of study design

The integrity of study design is evaluated with respect to treatment assignment, follow-up, and assessment of response. The primary interest is prospective studies. The problems with retrospective studies are briefly mentioned in the end.

Treatment assignment is critical for the integrity of a clinical study, because it determines the comparability of treatment groups at the beginning. Randomization is generally an effective technique to assign a large number of patients to a small number of treatment groups with

satisfactory comparability. Its effectiveness is, however, not guaranteed, especially when the number of patients is relatively small to the number of treatment groups. Therefore, the comparability of treatment groups needs to be carefully examined. A reliable method of examination is comparing baseline measures across treatment groups. Significant differences among treatment groups can seriously confound the effects of treatment.

Follow-up is essential to assessing the patients' responses over a time course. However, it also opens up opportunities for confounding. Concomitant treatments and patient withdrawals are two major sources of confounding. Lack of efficacy and toxicity are two common causes for the patients to seek concomitant treatments or withdraw from the study. Because their close relation to the effects of treatment, both concomitant treatments and patient withdrawals require early recognition, careful documentation and tight control. The key is to prevent them from occurring, provided that it is ethical, and to document the causes by all means if the events have occurred. Cause-specific comparison of concomitant treatments and patient withdrawals among treatment groups must be integrated in all well designed and executed clinical studies. Lack of such comparison should be considered a study flaw.

The assessment of response by both the patients and investigators may introduce bias toward favorite treatment for non-scientific purposes. Ideally, people who do not benefit from the study result should assess the responses. Blinding of treatment assignment during assessment of response has been a respectful practice in clinical research. If blinding is impossible, an independent third party may be hired. Both blinding and third party assessment are effective ways to exclude the potential bias from the knowledge of treatment and non-scientific interest. If neither blinding nor third party assessment is possible, the use of, what are generally believed, objective parameters is clearly a choice. An example is the so-called measurable disease in oncology trials. The last resort is the faith of investigators, declared by raising one hand in the air and putting down the other on the Bible. In front of suspicious and sophisticated audience, this presents the weakest defense against the accusation of assessment bias.

Retrospective studies are valuable for summarizing experiences accumulated over a long period of time. However, they generally do not

have the same degree of control as that attainable in prospective studies. Selection bias constantly troubles retrospective studies. It is difficult to find a group of people, as control, who are truly comparable to the selected series of patients. In addition, the history of the selected patients may be incomplete. Thus, any association between the current outcome and whatever a historical event may be entirely artificial. It is not surprising that most of the shocking conclusions made to the front page of newspapers are from retrospective studies. Comparing to prospective studies with comparable treatment groups and appropriate follow-up, the strength of evidence from retrospective studies is weak.

9.3.2 The sufficiency of observations

While a good design eliminates the confounding from structural flaws, sufficient quantity of observations diminishes the confounding from the uncontrolled factors. The quantity of observations directly determines the reliability and robustness of study results, and it is the single most important parameter in determining the strength of evidence. Without sufficient quantity of observations, nothing else matters.

No single criterion is universally accepted for determining the sufficiency of observations. Statistical power and the sample size determined with statistical power are fictions and have absolutely no role in determining sufficiency. Sensitivity, defined as the reciprocal of the number of patients, may be used for frequency endpoints, such as incidences of adverse effects, response rate, and etc. This measure is discussed in detail in Chapter Eight, section 8.9.2. Robustness, measured by the extent of fluctuation caused by small disturbances, may be used for quantitative measures.

A simple measure for assessing robustness is the change or percent change caused by deletion of few extreme data values. Suppose mean is the measure of interest. The change of mean upon deleting 2 extreme values is defined as

change(2-) = mean with all data – mean with two extreme values deleted,

where the two extreme values are either greater or less than the mean; in other words, they are on the same side of the mean, whichever more away from the mean. The percent change is defined as

$$\text{percent change(2-)} = \frac{\text{change (2-)}}{\text{mean with all data values}} .$$

If the extreme values excluded are moderately different from the mean, the percent change after data deletion measures the sensitivity of the mean to the impact of the uncontrolled factors. If the number of patients is small, deletion of few data values may dramatically change the mean; on the other hand, if the number of patients is large, the effect of data deletion is proportionally dampened, and the mean may not change at all. Because extreme values are mostly due to the effects of the uncontrolled factors, the sheer quantity of observations affords an excellent control of confounding.

9.3.3 Documentation in clinical study reports

Documents are the vehicle carrying intellectual products. By possession of documents, the society enjoys intellectual products without heavily relying on the individual producers. In this sense, documentation is an institutional behavior that holds together individuals and the institution to produce intellectual products, yet keeps both parties independent in a long run. Clinical studies are intellectual activities, and clinical study report is the final product. Therefore, the importance of documentation in study report cannot be overemphasized. This section lists some of the most important points that require careful documentation in order for readers of the study report to fully evaluate the study.

The intended patient population for the study needs to be clearly defined. A tabulation of eligibility criteria is usually enough. The actual patients may differ significantly from the intended patient population. Thus, a description of the actual patient population is necessary. The difference between the intended and actual patient populations also partially reflects the quality of study execution.

Treatment arms and controls must be clearly defined. There is little dispute in this regard. If an active control is used in the study, it would be

helpful, although not essential, to review its therapeutic profile in the literature. A significant deviation from what are commonly reported in the literature should trigger the alarm for unusual things in the study and the vigilance for its quality.

Treatment assignment must be carefully documented. Randomization remains the gold standard for treatment assignment. Nevertheless, the effectiveness of randomization is not guaranteed. Thus, the comparability of treatment groups must be thoroughly examined. Balanced distribution of baseline measures among treatment groups is convincing evidence on the effectiveness of randomization. Traditionally, randomization is cast *before* patient enrollment, and baseline measures are compared after the completion of data acquisition. With the advance of telecommunication technique, dynamic randomization will be more and more available for studies of moderate size for which simple randomization may not produce satisfactory comparability. In dynamic randomization, baseline measures are constantly monitored *during* patient enrollment.

Follow-up and timely assessment of responses must be documented in detail. Careful follow-up and unbiased timely assessment are as important as randomization for the control of confounding. They must not be overlooked in study report. The schedule and windows of follow-up need to be clearly defined, and compliance to schedule needs to be evaluated. Compliance is not only a parameter to assess the quality of study execution, but also an important means for the control of confounding from the time factor, especially when time is vital for the disease process and the effects of treatment.

Documentation of cause-specific patient withdrawals and concomitant treatments is part of the assessment of treatment effects. Without a cause-specific analysis of withdrawal and concomitant treatment, the study is inconclusive. A 15% withdrawal rate is expected in even a best designed and executed clinical study, and without documentation, we just cannot assume that those withdrawals are not the consequences of an inefficacious or toxic treatment. Similarly, unrestricted concomitant treatments can seriously jeopardize an otherwise perfect clinical study by rendering the final result uninterpretable. It is well possible that the patients were seeking concomitant treatments simply because the

treatment under study is inefficacious, too toxic or too cumbersome to administrate.

The actual patients included in analysis must be clearly defined. The so-called intent-to-treat (ITT) population consists of all the patients who are actually treated and whose responses are, at least, partially assessed. This is the patient population preferred by most researchers, because it represents the actual patients in the study. The so-called per-protocol (PP) population consists of patients who belong to the intent-to-treat population but also meet the eligibility criteria. This is the planned patient population interested by some for a what-if type of analysis. Analyses on subsets of patients, the so-called subset analyses, should be encouraged to explore heterogeneity. In fact, PP analysis, the analysis on the PP population, is a subset analysis. The multiplicity argument against subset analyses is baseless. The results of subset analyses can be equally valid and reliable if the comparisons are not confounded and the number of patients is sufficient.

Finally, the data actually used in analysis must be clearly defined, and any data manipulation must be explicitly documented. The traditional endpoint analysis takes data values at the end of study. To take into account the data from patients lost in follow-up, a variant of endpoint analysis is what is called last-observation-carried-forward (LOCF), where endpoint is defined as the last observation in study. Suppose patients are seen in clinic weekly up to twelve weeks. If a patient drops out at week three, by LOCF, this patient's data collected at his or her last visit, week three, will be put together with other patients' data collected at the end of week 12. Endpoint analysis ignores the effects of time. More problematic with endpoint analysis is that no one knows when is the end. A narrow focus on an arbitrary end point results in loss of information. This practice, once again, has much to do with the unrealistic statistical theory of Neyman and Pearson. Interested readers are referred to Chapter Ten on criticism of that theory. The most informative analysis is comparison of response profiles over the time of follow-up, what we call profile analysis. Profile analysis requires no data manipulation across different time points and fully utilizes all the data from study.

9.4 Graphical analysis of integrated studies

Graphical analysis of integrated studies is visualization of data or study results by their quality. The idea is not different from that in Chapter two, except that the studies are stratified by measures of their quality. If data are available from all the integrated studies, graphical techniques may be used to visualize them. If data are not available, and study results are expressed with means and their standard deviations for continuous data and count or percentage for categorical data, graphical techniques may be used to visualize their magnitudes. For the latter, information on data distribution is not available. Complex measures, as odds, odds ratio, relative risk, or any other transformations, will not be used to present study results. By graphical analysis, rich information is condensed in the visual field, and both consistency and heterogeneity are shown at the same time. When the number of studies to be integrated is large, graphical analysis is much more efficient than tabulation of numeric numbers.

9.4.1 Stratification by quality

The qualities of integrated studies are assessed with respect to integrity of design and quantity of observations. For the sake of discussion, study designs are graded into three categories: the good, the bad and the arguable. Good study designs share the following features: comparable treatment groups by effective randomization, unbiased assessment of responses, well planned follow-up schedules with good compliance, and patients withdrawals and concomitant treatments being unrelated to treatment. Bad designs are simply the opposite: incomparable treatment groups resulting from either ineffective randomization or non-randomized treatment assignment, potentially biased assessment of responses, as-needed follow-up or noncompliance to the scheduled follow-up, and lack of documentation on the causes of patient withdrawals and concomitant treatments. Table 9.1 on the next page portraits the good, the bad and the arguable.

Between the good and bad is the arguable. Evaluation of these studies is not quite straightforward. If treatment assignment is not randomized, even though treatment groups are comparable with respect to baseline measures, their comparability is still arguable because chances are that they are not comparable with respect to measures not included in baseline

measurement. As another example, if time is known not critical, then strict enforcement of follow-up schedule may not be necessary at all. On the other hand, if the actual visit times are comparable between treatment groups even though a regular follow-up schedule is not planned, the confounding from time may be negligible. If patient withdrawals and concomitant treatments are clearly related to treatment, they themselves are measures of treatment effects and must be incorporated into data analysis. However, if the frequency distributions of patient withdrawals and concomitant treatments are comparable among treatment groups even though the causes are not documented, one may argue that these events be not related to treatment since treatment makes no difference in these regards.

Table 9. 1 Assessment of Design: the Good, the Bad, the Arguable

	The Good	The Arguable	The Bad
Treatment assignment			
• randomized	Comparable		Not comparable
• systemic		Balanced	Not balanced
• by investigator		Balanced	Not balanced
Assessment of response			
• blinded/third party	All parameters		
• by investigator		Objective parameters	Subjective parameters
Follow-up			
• scheduled	Good compliance	Some compliance	Bad compliance
• as needed by patients		Equally often in groups	Not equally often
Withdrawals			
• cause documented	Not related to treatment	Related to treatment	
• cause not documented		Equally distributed	Not equally distributed
Concomitant treatments			
• cause documented	Not related to treatment	Related to treatment	
• cause not documented		Equally distributed	Not equally distributed

The quantity of observations is straightforward a parameter. Studies of similar design may be simply sorted by this parameter. One may group studies by their quantity of observations. The grouping can go by gut feelings or some criteria like the previously discussed sensitivity and robustness.

9.4.2 Graphical arrays

Lechat et al.* performed a meta-analysis of 18 double-blind, placebo-controlled randomized studies for the clinical effects of β blockade in the treatment of chronic heart failure due to either idiopathic dilated cardiomyopathy or coronary artery disease. The data are tabulated in Appendix E. The endpoint parameters are death rate, hospitalization rate and left ventricular ejection fraction. The studies are characterized with the numbers of patients in the placebo and treatment arms, drug names, and the duration of blind treatment. No information was given in the article for assessing the integrity of the designs.

The following graphs present the death rates. The studies are sorted by the number of patients, shown at the top of each bar chart:

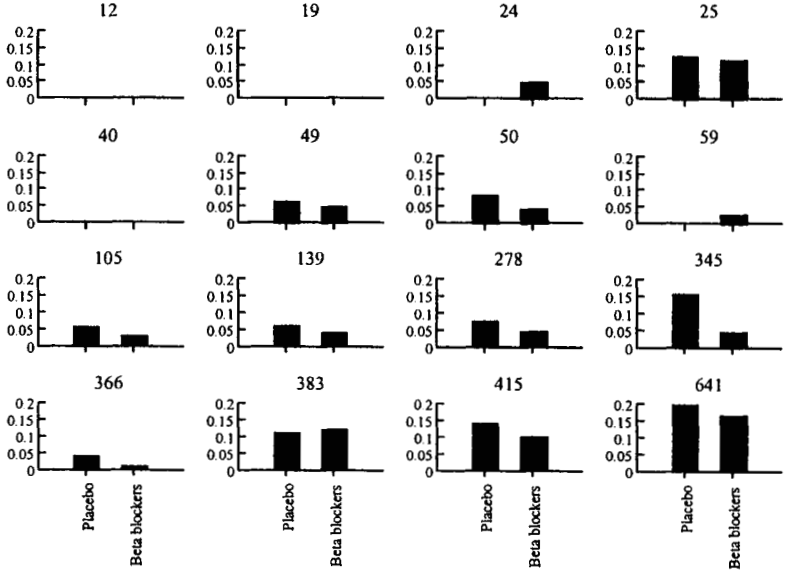


Figure 9.1 Death rates by treatment from studies of increasing sample size

*Circulation. 1998; 98:1184-1191. Copyright of American Heart Association, Inc. Used with permission.

Except for the study with 383 patients in the last row and second column, a lower mortality is consistently demonstrated in treatment groups. The magnitude of difference ranges roughly from 2% to 5%. The largest difference is observed in the study with 345 patients in the third row and last column, where 261 patients were assigned to treatment, and only 84 to placebo.

The following is a similar graph showing the mean left ventricular ejection fractions with their standard deviations:

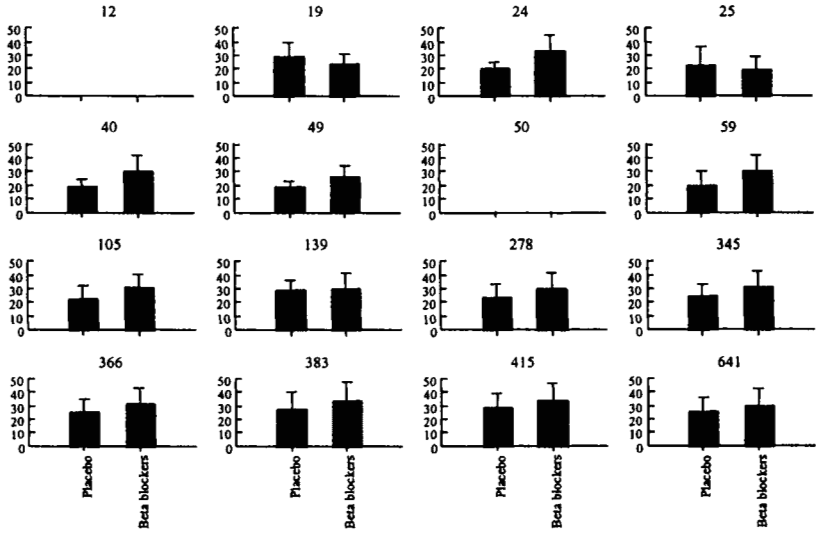


Figure 9.2 Left ventricular ejection fraction by treatment from studies of increasing sample size

A higher ejection fraction, by about 5% on average, is consistently demonstrated in most of the treatment groups. Inconsistency is demonstrated in the first row among small studies. Interestingly, when the sample size goes beyond 40 patients, little precision is gained for the mean ejection fraction with increasing numbers of patients. The standard deviations are about 30% of the corresponding means, and this is fairly consistent in studies shown in the last three rows.

To evaluate whether the duration of blind treatment and uneven allocation of patients between treatment and placebo have any impact on the results, the studies are partitioned with respect to these two factors. The following graph shows death rates by treatment, number of patients, duration of blind treatment, and the ratio of patients on placebo and treatment. The duration of blind treatment is divided into less or greater than ten months, and the ratio of patients on placebo and treatment into less or greater than 35%.

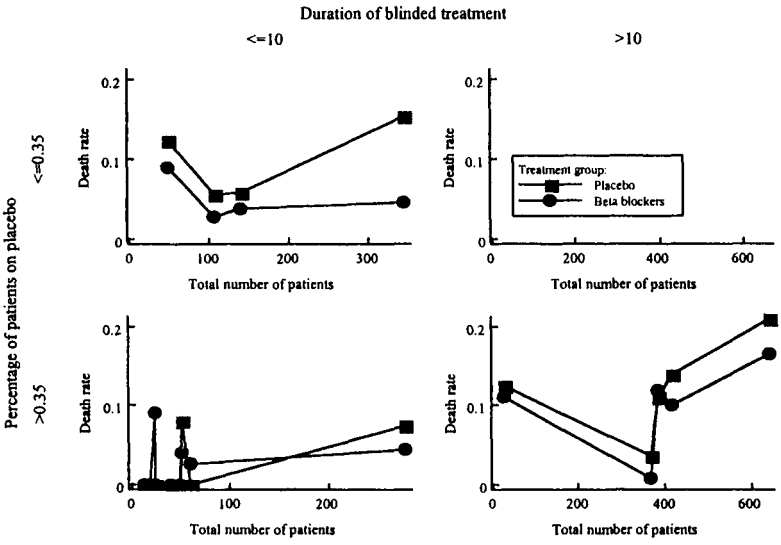


Figure 9.3 Death rate by treatment from studies of different sample size, duration of treatment, and patient allocation between treatment groups

A better consistency is demonstrated among long-term, large studies. Short-term, small studies present the greatest heterogeneity. There are four studies where less than 35% of the patients were assigned to placebo. This uneven distribution of patients raises the question of comparability. Suppose that group A has five patients and group B has ten. If only one patient died in each group, the mortality rates would be 20% for group A and 10% for group B, and their difference is 10% and their ratio is 200%. It is well possible that the difference between groups A and B is

insignificant, and the death is entirely caused by the uncontrolled factors. The difference in mortality rate is simply an artifact from using mortality as a measure in groups with different numbers of patients.

A similar graph shows the left ventricular ejection fractions:

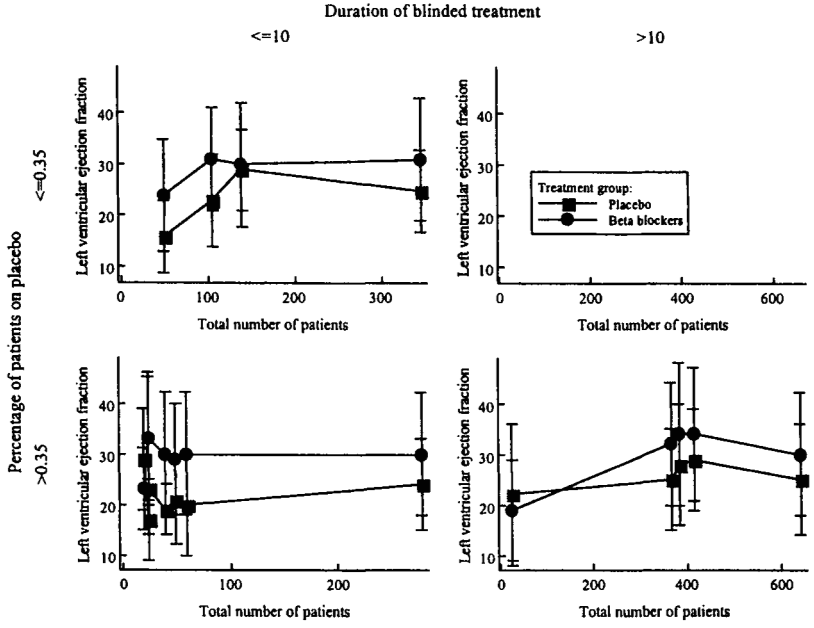


Figure 9.4 Left ventricular ejection fraction by treatment from studies of different sample size, duration of treatment, and patient allocation between treatment groups

The duration of blind treatment does not seem to have much impact on ejection fractions. Inconsistencies are mainly shown in small studies and, perhaps, studies of uneven patient distribution.

Graphical analysis can also be used to compare responses among subsets of patients. For illustration purpose, let us compare hospitalization rates between patients taking selective β_1 blockers and non-selective blockers. Again, the studies are categorized by the duration of blind treatment and total number of patients:

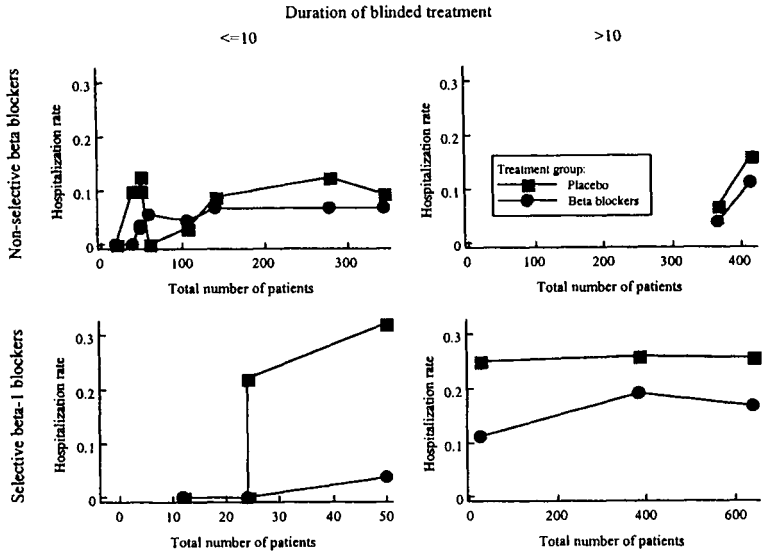


Figure 9.5 Hospitalization rate by treatment and selectivity from studies of different sample size and duration of treatment

For both selective and non-selective β blockers, lower admission rates are consistently observed in large studies. Inconsistencies are once again shown in short-term, small studies. In the four long-term, large studies in the second column of graphic array, selective β blockers apparently made a larger difference than did non-selective β blockers.

The above graphical analyses of integrated studies differ from the meta-analysis of Lechat, *et al* in several aspects. The main interest is to look for consistency and heterogeneity among the studies. The studies are stratified by several factors that potentially influence the results. A great deal of heterogeneity is simply explained by sorting the studies by the number of patients. Simple measures, as rate and the number of patients, are used to summarize the responses. Complex measures, as relative risk, odds or odds ratio, are not used. Agglomerate averages over the studies are not attempted. The studies are not differentiated by their primary endpoint parameters for the purpose of integration, although it is aware that the selection of primary endpoint parameters at the time might have

greatly influenced the design of individual studies. Statistical testing has no role in the analyses.

9.5 The analysis of variance on pooled data

The availability of data, not merely summary measures, from integrated studies presents a new opportunity for true scientific democracy. It allows for thorough assessment of study quality and detection of fraud. Data may be analyzed and appreciated for different purposes, from different perspectives, and with different expertise and sophistication. Nobody has to follow anybody else without personal experience with the data. Data, when made accessible to the public, ultimately benefit the society and can save tremendous resource from full utilization of information.

The aim of this section is to demonstrate how the analysis of variance technique can be utilized to analyze the data from integrated studies. The technique is convenient for computing averages or weighted averages in a variety of settings. The main purpose is to explore heterogeneity. A general analysis of variance technique is developed, based on the maximum likelihood technique of Ronald A. Fisher, to compile complex studies.

9.5.1 Common ground and appropriate stratification

Studies to be integrated may all be different. To do analysis of variance over different studies, the key is to identify the common ground to put together the studies and the cause of variations for appropriate stratification. This section discusses several common scenarios where integration of different studies can be performed with simple linear model techniques.

Studies with the same treatment groups in similar clinical settings may be pooled together. The following table illustrates two parallel studies,

Study 1	Placebo	Drug A
Study 2	Placebo	Drug A

The mean responses between treatment groups can be simply compared with the analysis of variance specified in the following linear model:

responses = study + treatment + residuals.

Treatment is the common factor, and its effects are measured with the means over the two studies. The difference between the two studies may account for significant variations of the responses, and it is represented by the effects of study and measured with two grand means. Here, the grand mean is the average of all data values in each study, regardless of treatment.

If the studies to be integrated are stratified, by center, prior treatment, and etc., the stratification should be included in analysis. The following table illustrates two parallel, multicenter studies:

Study 1	***Centers***	Placebo	Treatment A
Study 2	***Centers***	Placebo	Treatment A

The patients are stratified by center. An appropriate analysis is specified with the linear model:

responses = study + center(study) + treatment + residuals, or
responses = study + center(study) + treatment +
treatment-center(study) interaction.

In these models, center(study) denotes center nested in study, meaning that patients are grouped by studies and then sub-grouped by centers within each study. In this analysis, the effects of center are taken into account. By grouping centers into study, the effects of study specified in the linear model are measured with the sum of center effects within each study.

Even studies with different stratification can be put together. The following table illustrates three parallel studies with different stratifications:

Study	*Strata*	Treatment Groups	
Study 1	***Centers***	Placebo	Treatment A
Study 2	***History***	Placebo	Treatment A
Study 3	***None***	Placebo	Treatment A

Study 1 is stratified by center, study 2 history, and study 3 is not stratified. Since the studies are stratified differently, when they are pooled together, generic term "strata" is used to represent the strata generated with different stratifications. The analysis is either

responses = strata + treatment + residuals, or
responses = strata + treatment + treatment-strata interaction.

Study 3, as a whole, is viewed as a stratum. The effects of study are not included in the analysis. Strata are nested in studies, and for study 3, study and strata are synonyms.

Studies of the same treatment in different patient populations may be put together. Suppose that the same design is carried out in adult and pediatric patients, as illustrated in the following table:

Study 1	***Adults***	Placebo	Treatment A
Study 2	***Children***	Placebo	Treatment A

Appropriate analysis should include the interaction effects of study and treatment:

responses = study + treatment + study-treatment interaction + residuals.

A significant study-treatment interaction effect implies that the effects of treatment are different in two patient populations. Significant treatment effects, in the absence of study-treatment interaction effects, imply that the mean responses, over both patient populations, are different between treatment groups. But this difference can result from, say, $10 - 5 = 5$ or $20 - 15 = 5$. The effects of study indicate the absolute magnitudes of mean responses in each patient population.

In general, studies of different treatments should not go together. An exception is when treatments differ only in quantity. Dose escalating studies are a typical example:

Study 1	Placebo	Drug A at 5 mg bid	Drug A at 20 mg bid
Study 2	Placebo	Drug A at 25 mg bid	Drug A at 40 mg bid

If we are interested in dose-specific mean responses, the two studies cannot be combined. The effects of Drug A at 5 mg bid in study 1 are not directly comparable to that of Drug A at 25 mg bid in study 2. If, however, the primary interest is dose-response relationship, measured with mean response curves, the two studies share a common ground, and they can be simply combined on that ground. A linear model for the analysis is this:

responses = study + treatment + study-treatment interaction + residuals,

where treatment is viewed as a continuous variable indicating doses, and it represents a linear dose-response relationship. Since there are three dose levels in each study, a quadratic dose-response relationship could also be specified. Significant study-treatment interaction means that the mean dose-response curves in two studies are not parallel. Significant treatment effects, in the absence of study-treatment interaction effects, suggest that the mean dose-response curves are not flat. Significant study effects indicate that the mean dose-response curves separate by some distance between two studies.

If the interest is response profiles over time, we need to pay attention to follow-up schedules. Studies may be pooled when the follow-up schedules are similar. The following table presents two studies with similar follow-up schedules:

	Visit Window: Week ± days				
Study 1	1 ± 3	2 ± 3	5 ± 7	10 ± 7	20 ± 14
Study 2	1 ± 3	2 ± 3	5 ± 10	10 ± 14	20 ± 20

These two studies can be directly combined. The week numbers can be used as chronological marker, and comparisons can be made at each visit.

Studies with different visit schedules present a problem. The following table illustrates two studies with different follow-up schedules:

	Visit Window: Days ± 5				
Study 1	7	14	28	42	70
Study 2	7	21	35	49	63

This lack of common chronological marker renders it impossible to pool these two studies for making comparisons in the same time frame. However, if the primary interest is the mean response curves over time, not the mean responses at specific time points, the two studies may be put together with the same technique for comparing dose-response curves. The analysis may be specified with this linear model:

$$\text{responses} = \text{study} + \text{treatment} + \text{poly(time, 2)} + \text{treatment-poly(time, 2) interaction},$$

where poly(time, 2) denotes quadratic curves. The term, treatment-poly(time, 2) interaction, measures the difference of mean response curves among treatment groups. If we expect different response profiles in different studies, we may add interactions with study:

$$\text{responses} = \text{study} + \text{treatment} + \text{poly(time, 2)} + \text{treatment-poly(time, 2) interaction} + \text{study-}\{\text{treatment} + \text{poly(time, 2)} + \text{treatment-poly(time, 2)}\}\ \text{interaction} + \text{residuals}.$$

The last interaction is the sum of following interaction effects: study-treatment, study-poly(time, 2), and study-treatment-poly(time, 2). A significant effect from any of these interactions indicates some differences of mean response profiles between studies.

By focusing on mean response profiles, represented with curves, as opposed to specific mean responses at points, we gain technical advantages. We may appreciate this with our geometric experience that two points make a line, and three points could make a curve. If we assume that the response profiles can be represented by smooth curves, then few points may be all it takes to estimate those curves.

9.5.2 An analysis of variance technique for compiling studies

Most of the analyses on multiple studies can be simply conducted with a linear model and clever parameterization. The technique developed here is for integration of complex studies to which a single linear model cannot accommodate. The idea is to use two surrogate measures to compile complex studies, and it comes directly from the maximum likelihood technique of Ronald A. Fisher.

Factors in complex studies are partitioned into study-specific factors and the common factors. The table below illustrates this partition:

	Study-specific factors	**The common factor**
Study 1	Design factors, covariates, denoted by β_1	Treatment, denoted by τ
Study 2	Design factors, covariates, denoted by β_2	Treatment, denoted by τ

Study-specific factors are specific to the study, not shared by other studies to be integrated, and they are local factors. The common factors are those shared by all the studies, and they are global factors. The mean responses measuring the effects of study-specific factors are estimated from the study itself; the mean responses measuring the effects of common factors are estimated from all the studies. The following linear models represent the partition:

Study 1: responses = study-specific factors (β_1) + treatment (τ) + residuals,

Study 2: responses = study-specific factors (β_2) + treatment (τ) + residuals.

The maximum likelihood technique is used to compute the mean responses for both study-specific factors and the common factors. In essence, the contribution from each patient to the mean response is quantified with an efficient score, and its quality is measured with the Fisher information. Efficient score is an intermediate quantity in the maximum likelihood technique. For each parameter in the linear model, its efficient scores represent the contributions from individual patients to the mean response represented by that parameter. For the effects of study-specific factors, the mean responses are estimated from the efficient scores of that specific study. For the effects of common factor, which is

treatment in the above linear models, the mean responses are estimated from the sum of efficient scores in all the studies. The Fisher information associated with each efficient score can be simply added up to measure the quality of mean responses. Appendix C details the mathematical development of this technique.

For complex studies, the advantage of efficient scores over individual data values is threefold. First, the effects of common factors are more precisely estimated by adjusting for the effects of study-specific factors. Second, the efficient scores can be simply added up to compute the mean responses across any studies or strata. This flexibility is particular appealing to integration of complex studies. Finally, the efficient scores themselves can be viewed as individual data values for assessment of heterogeneity. A graphical display of efficient scores along with the suspected causes of heterogeneity is perhaps all we have to do.

9.6 Some other techniques in meta-analysis

The analysis of variance on summary measures essentially characterizes the technical development of meta-analysis. However, the main focus of meta-analysis, at least presently, has been statistical hypothesis testing, which, in my opinion, does not help improve our understanding of the data. Agglomerate averages over heterogeneous studies have no practical meaning. It is probably more scientifically profitable focusing on the heterogeneity of summary measures. This section demonstrates how the analysis of variance technique may be utilized to explore heterogeneity, criticizes the measure of effect size and clarifies the concept of random effects.

9.6.1 The analysis of variance on summary measures

The idea and technique are exactly the same as those discussed in Chapter Four except that, first, the data are consisted of summary measures, and second, measures on study quality are incorporated into the analysis for explanation of heterogeneity. The data published by Lechat *et. al.*, tabulated in Appendix E and analyzed in section 9.4.2, are re-analyzed to illustrate the technique. The endpoint parameters are death rate, hospitalization rate and left ventricular ejection fraction. The independent variables are treatment (placebo or a β blocker), study

identification, and the duration of blind treatment. Due to limited quantity of summary measures, the comparison of treatment effects is stratified with respect to only one independent variable at a time.

The following linear models specify the effects of treatment stratified by study:

death rates, hospitalization rates, or left ventricular ejection fractions = study + treatment + residuals.

The results are summarized in the following ANOVA table:

Response variable	Source of Variation	DF	Sum of Squares	Mean SS	Ratio	P-value	LS Means β/Placebo
Death rate	Study	17	0.1044	0.0061	7.90	0.0001	0.0571
	Treatment	1	0.0019	0.0019	2.42	0.1379	0.0652
	Residual	17	0.0132	0.0008			
Hospital rate	Study	17	0.1738	0.0102	2.96	0.0157	0.0571
	Treatment	1	0.0378	0.0378	10.94	0.0042	0.1219
	Residual	17	0.0588	0.0035			
LVEF	Study	15	307.00	20.467	1.46	0.2375	29.563
	Treatment	1	300.12	300.12	21.35	0.0003	23.438
	Residual	15	210.88	14.058			
LVEF	Study	15	35.096	2.3397	1.66	0.1673	29.345
Weighted by	Treatment	1	29.373	29.373	20.89	0.0004	23.182
1/STD	Residual	15	21.091	1.4061			
DF: degree of freedom, SS: sum of squares, β/Placebo: LS means of β blocker group, above, LS: least square, STD: standard deviation. and placebo group, below.							

Strong study effects suggest large variations among the study results. Both death rate and hospitalization rate are quite heterogeneous from study to study. For left ventricular ejection fraction (LVEF), however, variations among studies are rather small. Considerable treatment effects are demonstrated on hospitalization rate and LVEF. The treatment effects on death rate are, however, very small. Weighted analysis with the standard deviations of the means does not change the result on LVEF.

To explore the cause of heterogeneity on death rate and hospitalization rate, the comparison of treatment effects is stratified with respect to selectivity and duration of blind treatment. Selectivity is 1 if the β blocker

is selective on β_1 receptors, and 0 otherwise. The analyses are specified with the linear models:

death rates, hospitalization rates =
selectivity + duration + treatment + residuals.

The results are summarized in the following table:

Response Variable	Source of Variation	DF	Sum of Squares	Mean SS	Ratio	P-value	LS Means
Death rate	Selectivity	1	0.0058	0.0058	2.58	0.1179	0.0478 β_{12}
	Duration	1	0.3931	0.3931	17.58	0.0002	0.0738 β_1
	Treatment	1	0.0019	0.0019	20.84	0.3656	0.0536 D
	Residual	32	0.0716	0.0022			0.0681 P
Hospital rate	Selectivity	1	0.0346	0.0346	7.97	0.0081	0.0647 β_{12}
	Duration	1	0.0566	0.0566	13.07	0.0010	0.1283 β_1
	Treatment	1	0.0378	0.0378	8.73	0.0058	0.0641 D
	Residual	32	0.1387	0.0043			0.1289 P
DF: degree of freedom, SS: sum of squares, LS: least square, STD: standard deviation.				D: β blocker, P: placebo β_{12}: Non-selective, β_1: selective β blocker			

It seems that the duration of blind treatment is an important cause of heterogeneity in both mortality and hospitalization. The analysis shows that β blockade significantly reduces the chance of hospitalization. Less hospitalization is seen in studies with non-selective β blockers. Notice that selectivity refers to study, not treatment. In studies with non-selective β blockers, the patients may take either placebo or a non-selective β blocker. Therefore, non-selective β blockers do not necessarily cause less hospitalization seen in *studies* with non-selective β blockers. Here, inclusion of selectivity in the analysis is just a demonstration how to explain heterogeneity with the analysis of variance technique.

Because little data values are available, the analysis of variance on summary measures is limited to studies of very few factors at a time. Compared to graphical analysis, the analysis of variance allows us to calculate the mean responses in different categories and the trends over continuous variables. However, graphical analysis presents complete data, and the data can be evaluated in broader perspectives. The analysis of variance is an effective technique, but should never substitute for data visualization.

9.6.2 Effect size versus group-specific measures

Effect size is an early measure in meta-analysis proposed for compiling clinical studies. The idea is to use a single measure to summarize the result of a study. For instance, if the response rates to treatment A and placebo P in a study are reported to be 80% and 40%, respectively, the effect size can be either D = 80% – 40% = 40%, R = 80% ÷ 40% = 2, or odds ratio = (80% ÷ 20%) ÷ (40% ÷ 60%) = 6. If the response variable is continuous, and the mean responses to treatment A and placebo P are 10 and 20, and their corresponding standard errors are 5 and 4, a definition of effect size is

$$\text{Effect size} = \frac{\text{mean}_{\text{treatment A}} - \text{mean}_{\text{placebo}}}{\text{standard error of placebo}} = \frac{10-20}{4} = -2.5.$$

Another definition is essentially the same except that the pooled standard error, instead of the standard error of placebo, is used in the denominator.

Effect size is an awkward measure and suffers from serious loss of information. If a study has three treatment groups, not only is there a lack of unique definition of effect size, but also more than one effect sizes have to be defined to represent the information in three treatment groups. However, if multiple effect sizes are used to summarize a study, the original idea of parsimony can no longer be fulfilled. Even if a study has only two treatment groups, an effect size of whatever definition cannot fully preserve the original information in group-specific measures. The following table, for instance, well presents the fact that 240 out of 6500 high risk women taking placebo developed breast cancer while 120 out of 6500 taking tamoxifen developed breast cancer over six years:

Exposure	Breast Cancer	Total Patients	Risk
Tamoxifen	120	6500	2%
Placebo	240	6500	4%

If the result is expressed in terms of following effect sizes, it can be seriously misleading. The first effect size is relative risk, which is 2 = 4% ÷ 2%; the other is risk reduction, which is –50% = (2% – 4%) ÷ 4%.

A 50% risk reduction does not tell us whether it is 50% versus 25% or 0.2% versus 0.1%. There is no doubt that a 50% risk reduction is news making and sufficient to justify the risks of tamoxifen for prophylaxis. But if we look at the actual incidences, the balance of benefit-risk calculation could have tilted the other way.

When data from published studies are not available, summary measures in each treatment group, not effect size or other derivations from them, should be the original data for meta-analysis. The most informative summary measures are the mean, standard deviation, number of patients, and number of events in each treatment group.

9.6.3 Variations among studies and the random effects

In meta-analysis, the variance of combined effect size over studies has two components: the variance within individual studies and the variance between individual studies. The within-study variance represents the sum of variations within each study, and the between-study variance represents the heterogeneity among studies. Suppose 10 studies each is summarized with an effect size and its standard error. Most of the combined effect size in meta-analysis, such as combined odds ratio, is a weighted average:

$$\text{combined effect size} = \frac{\text{sum (effect size x weight)}}{\text{sum (weights)}}.$$

The most commonly used weight is the standard error of effect size from each study. By the conventional mathematical definition of variance, the variance of this combined effect size is

$$\text{variance of combined effect size} = \frac{1}{\text{sum (weights)}}.$$

This variance represents only the within-study variations and overlooks the between-study variations.

The overall variance of combined effect size should include variations both within and between studies, and its estimate should base

on the deviations between the combined effect size and contributing effect sizes:

$$\text{overall variance} = \text{mean (effect sizes in individual studies} - \text{combined effect size)}^2.$$

The difference between this overall variance and within-study variance measures the between-study variance:

$$\text{between-study variance} = \text{overall variance} - \text{within-study variance}.$$

Between-study variance can be negative if the within-study variance is greater than the overall variance. A negative between-study variance suggests that the effect sizes of poor precision are close in magnitude from study to study.

Because the conventional mathematical definition of variance fails to take into account between-study variation, the concept of random effects was introduced. This concept has generated a good deal of confusion, mainly due to its verbatim that the studies to be integrated are a "random" sample of all possible studies. This verbatim of "random" came from mathematical statisticians who grossly misunderstood the original concept of Charles Henderson. Chapter Six, section 6.5, has more discussions on random effects. "A random sample of all possible studies" is entirely an inoperable concept. The technical nature of random effects is to direct the mathematical manipulations so that the between-study variance is included in the calculation. If the primary interest is mean responses, the result from meta-analysis with random effects should not be capriciously different from that of the analysis of variance with an equivalent fixed-effects linear model.

This Page Intentionally Left Blank

10
The Fiction behind the Current Statistics and Its Consequences

Summary

This chapter criticizes the statistical theory of Neyman and Pearson and points out how the concept of errors and test of hypothesis in formal logic fostered by their theory most adversely impact the practice of clinical research. The issue of multiplicity raised from the concept of error discourages researchers from making as many observations and evaluations as possible. The rigid test of hypothesis with formal logic discourages the use of active control and makes it impossible to draw conclusions from studies designed for showing equivalence. P-value and confidence interval are interpreted as measures, and the confusion surrounding them is clarified. The absurdity of power and determination of sample size with power is demonstrated. Finally, the maximum likelihood technique is introduced, and the technical nature of mathematical distribution is explained.

10.1 An overview

The discipline of statistics deserves credit for providing graduates with

basic quantification skills essential to operating in the demanding research society. However, some foundation of that discipline is not compatible, at least in my opinion, with the principles of learning and experimentation cherished by most clinical researchers. Graduates from that discipline are more or less influenced by that foundation, and they carry that influence into their comments and recommendations on research proposals, publications and government regulations, which in turn contribute to the decisions on distribution of research funding and permission of marketing for profit.

In recognition of the difference between the statistical theories and the principles of clinical research, medical researchers need to have a realistic image of statistics as a discipline. A realistic image is not only important to working effectively with statisticians trained in the current statistical educational programs, but also crucial to the right use of research methodology, without being confused by those unrealistic statistical theories. The aim of this chapter is to discuss the fundamental concepts and techniques in statistics that have profound impact on the design and evaluation of clinical studies.

10.2 The fantasy of truth and the concept of errors

That statistics can tell the truth or at least indicate how close a sample is to the truth is the ultimate fantasy held consciously or subconsciously by most of the statisticians trained in the current statistical educational programs. The paradigm adopted in the current statistical academia is largely founded on the concept of errors, entirely due to Jersey Neyman and Egon Pearson. In a collection entitled *Joint Statistical Papers*, Neyman and Pearson argued that people make two kinds of error when making a decision on rejection or acceptance of a hypothesis:

- the error of rejecting a hypothesis that is in fact right — the error of the first kind or the type I error, and
- the error of accepting a hypothesis that is in fact wrong — the error of the second kind or the type II error.

This argument is conveniently summarized in Table 10.1 on the following page. The decision theory of Neyman and Pearson states that a decision-making process can be simply formulated into a mathematical process that minimizes the probability of making both types of errors. The probability of type I error is known as the p-value, denoted by α, and the complement

of p-value, $1 - \alpha$, is known as the confidence level; the probability of type II error is denoted by β, and its complement, $1 - \beta$, is known as the power.

Table 10.1 Decision-Making Theory of Neyman and Pearson

	DECISION	
TRUTH:	**Rejection of hypothesis**	**Acceptance of hypothesis**
Hypothesis is right	Type I error: Reject the right hypothesis	Confidence level: Accept the right hypothesis
Hypothesis is wrong	Power: Reject the wrong hypothesis	Type II error: Accept the wrong hypothesis

The concept of errors is unrealistic. The fundamental problem is that we do not know the truth. The reality of clinical research is that we learn by experimentation and hope to know the truth, not the other way around that the truth is already known and is available for judging our observations. A decision based on the currently available information may prove to be wrong only when further information is available. When such information is not available, it is logically impossible to evaluate a decision with the current information from which the decision was made. For a single study, the only available information comes from the study itself. A decision to reject or accept a hypothesis can only be made after scrutiny of available information and careful balancing of conflicting interests. Until information from subsequent studies becomes available, there is no information to evaluate whether the current decision is right or wrong. Therefore, the concept of errors has no meaning in any sense for a single study.

10.3 Assumption of distribution

Like all mathematical deductions, the theory of Neyman and Pearson is essentially a formal logic. The axiom or precondition of their theory is knowing the truth. Because the true is uncertain or unknown, they must come up with a solution to sustain their theory. By using mathematical distribution to symbolize the unknown truth, Neyman and Pearson cleverly chose an entity to represent uncertainty or ignorance and passed that ignorance to an uncommitted conclusion, known as the probability statement. By symbolizing the truth, they avoided the burden of knowing

the truth, and by using an entity of unknown meaning to human inductive thinking, they met the requirement of formal logic.

A mathematical distribution is merely a mathematical function whose domain is limited from 0 to 1. A mathematical distribution itself has no information other than rules of computation. The following mathematical function is the popular normal distribution:

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}.$$

It is simply a series of calculations conveniently expressed with a set of Roman and Greek letters. When x, μ, and σ are actual numeric values, the calculations give rise to a numeric number between 0 and 1. Interestingly, the lack of a symbolic system to effectively document complex computations has been considered to be a major drawback of the Chinese language.

The use of mathematical distribution in statistics probably originates from the historical link between statistics and probability. Probability is a quantification of uncertainty or degree of ignorance. Gambling was the driving force behind the development of probability theory for calculation of uncertainty or risk. The development of mathematics and physics in the nineteenth century created a paradigm of deductive thinking that was once pushed to the point that our knowledge of the world was believed to be self-contained and could be entirely deducted. Although attempts to prove self-containing, like all the machines designed for eternal motion without outside energy, ended up with total failure, the paradigm of deductive thinking is still embraced by many mathematicians who enjoy building a complex theory from few axioms. The advance of experimental science, however, demonstrated the power of yet another human reasoning, the inductive thinking, with the most renowned success being Charles Darwin's theory of evolution. In contrast to deductive thinking, it was well recognized that inductive thinking is not precise and carries uncertainty, and skepticism to the merit of inductive thinking was widely spread. Thanks to the experimental psychology pioneered by Jean Piaget, we now have a better understanding of inductive thinking as the link between concepts and operation. We have accepted inductive thinking as

the human intellectual behavior and pay our attention only to observed evidence. At the turn of the nineteenth to twentieth century, however, there were constant attempts to fit inductive thinking into the paradigm of deductive thinking by imposing elements of preciseness and formal logic. Quantification of the uncertainty in inductive thinking was one of those attempts. The probability theory, developed for gambling, was borrowed to cast the process of inductive thinking, a complex intellectual activity of our brain, into a mathematical framework. The statistical decision-making theory of Neyman and Pearson is an elaborate combination of formal logic and probability.

Because mathematical distribution is essential to the decision-making theory of Neyman and Pearson, statistical testing fostered from their theory always requires that the frequency distribution of real data resemble a mathematical distribution. The assumption of distribution in trial protocols is the commonly seen statement of this requirement. Because of its logical position in statistical testing, lack of verification of that assumption creates anxiety. For those who swallowed the theory of Neyman and Pearson, it is not irrational to verify the assumption of distribution. Unfortunately, no guidelines are available for verifying assumption of distribution. Data visualization has been widely used for this purpose. Trouble is to use statistical testing to verify assumption of distribution, which creates a logical circle because verifying an assumption requires the making of another assumption. As long as the statistical theory of Neyman and Pearson is used for decision-making, it is inevitable to make assumption of distribution, and the trouble of verification will go on and on.

In practice, the popular use of normal distribution partially eases the pain from making distribution assumption of unknown practical meaning. When the interest is the mean and its standard error, the assumption of normal distribution leads to a sensible comparison of their magnitudes, the mean ÷ its standard error, through the system of Neyman and Pearson. Such sensible comparison is generally not achievable with assumption of other distributions. Most of the time, the mathematical manipulation involved is so complex that construction of a sensible comparison relies solely on purposeful approximations. The Taylor series expansion is the commonly used approximation technique, justified with the asymptotic argument. The essence of asymptotic argument is that large sample size

justifies the first-order Taylor series approximation. The entire purpose of picking only the first-order Taylor series is directing the mathematical manipulation under the theory of Neyman and Pearson to a comparison between the mean and its standard error.

10.4 P-value

P-value is a reminiscence of the glorious time when humans were struggling to master their intellectual activities precisely with mechanics and mathematics. The persistent use of p-value in current statistical practice comes mainly from the stubborn embrace of the unrealistic formalism of Neyman and Pearson and the disrespect of the advance of experimental psychology on human cognitive behavior. Nevertheless, to be respectful to that glorious history on exploration of human reasoning power, p-value may still be used as it has been, but it must not be interpreted as a probability of error in the sense of Neyman and Pearson. P-value must be viewed as a measure expressing how reluctantly we would like to accept an opinion. Technically, p-value is meaningful only in the context of the analysis where it is defined and computed. An isolated p-value has no meaning, and p-values from different analyses are not necessarily comparable. Moreover, p-value does not follow any rules of probability calculation. Any mathematical manipulation of p-values by the theory of probability, such as the Bonferroni's protocol, is doomed to meaningless result.

For all practical purposes, p-value may be viewed as an equivalent measure to the comparison between measures of claimed effects and the effects of the uncontrolled factors. The p-value of permutation test measures the difference between what has been observed and what could have been observed had the patients been affected only by the uncontrolled factors. The p-value in the analysis of variance is equivalent to the comparison of mean to its standard error. The result of that comparison is matched to the standardized normal, or its equivalent t or F, distribution in order to obtain a p-value. The standardized normal, t and F distributions are mathematical distributions derived from the assumption of normal distribution. Matching to these distributions is nothing more than a change of scale. The relationship established through the matching is that the smaller the p-value, the better the precision of the mean.

As a measure, p-value does not improve our understanding of data in any way. Comparing to other measures, it carries the least amount of information. Choosing p-value as opposed to any other measures is, at most, like choosing the yard as opposed to meter to measure distance. There is no logical basis whatsoever to endorse the rule of p-value ≤ 0.05 for declaring significance.

10.5 Confidence interval

Confidence interval is equivalent to p-value but more involved with the underlying analytic procedure. For instance, the ratio of mean and its standard error being greater than 1.96,

$$\left|\frac{\text{mean}}{\text{standard error}}\right| > 1.96$$

gives rise to p-value of 0.05 when it is compared to the standardized normal or t distribution. Equivalent but complementary to that inequity is the ratio being equal to or smaller than 1.96,

$$\left|\frac{\text{mean}}{\text{standard error}}\right| \leq 1.96\,,$$

which can be rearranged directly into the 95% confidence interval:

$$\text{mean} - 1.96 \times \text{standard error} \leq \text{expected mean} \leq \text{mean} + 1.96 \times \text{standard error}.$$

In statistical testing, p-value and confidence interval are used to make probability statement of a statistical test result. For the ratio of mean and its standard error, the p-value of 5% means that the probability is 5% for the ratio to be larger than 1.96. The equivalent statement in terms of confidence interval is that the probability is 95% for the mean to be in the confidence interval. In reality, the probability of 5% does not imply that if the same study is repeated 100 times, 5 out of 100 times the ratios will be greater than 1.96. For a single study, this probability of 5% is meaningful only in the sense that the mean of 5% p-value is more precise than the mean of, say, 15% p-value. Similarly, the probability of 95% does not imply that if the same study is repeated 100 times, 95 out of 100 times the

means will be within that confidence interval. For a single study, the confidence interval of 95% is narrower than that of, say, 90%. It must be kept in mind that for a single study, whatever probability statement, no matter it is in terms of p-value or confidence interval, is an expression of attitude only, not a prediction of anything beyond that single study.

10.6 Devastating impact on clinical research

The concept of errors has devastating impact on the current clinical research practice. Because there is a chance of error from any test of hypothesis, a conclusion from multiple tests is inevitably opt to more chance of error than that from a single test. This creates a dilemma between the scientific need of evaluating as much information as possible and the resulting unfavorable p-value from multiple tests. This is the notorious issue of multiplicity. The concept of errors directly leads to the notion of statistical power and the practice of determining sample size with statistical power. Because statistical power is unrealistic and intangible with observable measures, it leads researchers astray from observable evidence to an endless logical gyrate. This is the mysterious issue of statistical power and sample size determination.

10.6.1 Multiplicity

With the statistical decision-making theory of Neyman and Pearson, multiplicity is a ubiquitous issue. This issue arises whenever the results of multiple statistical tests are consolidated for making a claim. Suppose that the p-values from two statistical tests are 0.05. By the theory of Neyman and Pearson, the p-value of 5% is the probability of making a wrong conclusion. If we draw a conclusion from the results of these two tests, the conclusion is right only if the results of both tests are right. In other words, the conclusion is wrong if the result of any of the tests is wrong. By the rules of probability, the probability of making a false claim from two tests each having a chance of 5% is 0.05 + 0.05 = 10%. Therefore, a conclusion from studying two parameters is more likely to be wrong than that from studying only one! In a system where p-values from statistical tests are heavily weighted in decision-making, the multiplicity issue creates a dilemma between the scientific need of evaluating as many parameters as possible and the disadvantage from the resulting

unfavorable p-value for making a claim. The more you study, the more the conclusion is likely to be wrong!

Technical strategies on statistical tests have been developed to get around this dilemma, and they are true mathematical nightmares. In statistical academia, the issue of multiplicity is also known as the problem of multiple testings or multiple comparisons. Mathematical statisticians attack this problem by either using order statistics to reduce the number of tests, or assuming a uniform distribution of the p-values and comparing the combined p-value to the chi-square distribution with one degree of freedom. Applied statisticians tend to favor a hierarchical pyramid for the parameters of interest. In research protocols, this pyramid often appears with a single "primary" parameter on the top followed by numerous "secondary" and even "tertiary" parameters. The advantage of testing only the primary parameter is often offset by the excessive risk of betting the entire trial on a single cast. The most bizarre strategy is using composite parameters, such as a linear combination of spirometric measures and symptom scores. The advantage of testing a giant combo is offset by the obvious absurdity.

Strategies of designing large "confirmatory" trials as opposed to small trials of good quality have to do with the issue of multiplicity. Multiple trials inevitably involve multiple testings, whereas a single trial eliminates this multiplicity. Researchers who run multiple trials to demonstrate consistency are punished with a diminished chance of making a successful claim. The discouragement of interim analyses when partial data become available from trial is a direct consequence of the multiplicity issue. This leaves researchers little opportunities to learn and improve clinical studies from immediate feedback. End point analysis as opposed to profile analysis over the time course is another strategy to avoid multiple testings. The result is unnecessary data manipulations and loss of information.

The issue of multiplicity unnecessarily aggravates the conflict between clinical researchers and regulatory authorities in the evaluation of clinical studies. Because there are so many ways to get a p-value, researchers might try everything to get a small p-value, and knowing this, reviewers are very skeptical of any analysis and reporting from the researchers. A documented analysis and reporting plan is often required before the researchers actually see the data. When data become available and the

planned analyses are found inadequate, any "ad hoc analysis" needs to be carefully explained in order to eliminate the suspicion that only results in favor of the researchers have been reported.

Ironically, some physicians promote the folklore that a normal patient is a less-tested patient and attribute this phenomenon to the problem of multiplicity. It is true that the more we test a patient, the more we will likely find an abnormal test result for the patient. However, a patient with an abnormal test result is not necessarily a sick patient. An abnormal test result could simply mean that the patient is an individual who is different from the patient population on which the norm is defined. While a test result labeled abnormal can be psychologically stressful for the patient and generate a chain reaction calling for more tests, there is no logical problem with collecting as much information as possible from a patient. Only ethics and resources could hold physicians from exhaustive investigation.

10.6.2 Statistical power and sample size

In the theory of Neyman and Pearson, power is the probability of correctly rejecting a wrong hypothesis. Because there is no knowledge *a priori* to judge whether a decision to reject a hypothesis is correct or not, the power is, again, an unrealistic concept with *no* practical meaning. To compute power, it is essential to assume not only that the frequency distribution of the data yet to be collected resemble a mathematical distribution, but also that a "clinically significant difference" yet to be found or confirmed in the study must be declared before the study. Defense for these assumption and declaration generally requires making other assumptions.

If the standard error is used in statistical test, the process for power calculation can be reversed to compute the sample size. The following statement, frequently appearing in clinical trial protocols, exemplifies the typical requirement for computing sample size. Assuming that the data follow a normal distribution and the standard deviation be 20, to detect a clinically significant difference of 20 with 5% error protection (type I error), at least 76 patients with 38 to each group are required to achieve a 99% power. The underlying statistical test for this statement is a t-test, the analysis of variance for comparing two groups. The declared "clinically significant difference" and the assumed standard deviation by and large

dictate the resulting power and sample size. A large difference and a relatively small standard deviation could result in a sample size as small as one patient. On the other hand, a small difference and a relatively large standard deviation could raise the sample size to as large as ten million patients. For instance, if we declare in the previous statement that the difference be 300, only four patients are required to achieve that magnificent 99% power, whereas 7356 patients would have to be recruited if the difference was declared to be 2.

Because power and the sample size determined with power are based on non-observable measures, they have never been truly useful for clinical research. Behind an apparently rigorous statement are negotiations on meaningless assumptions and declarations. The final power or sample size is merely an arbitrary number tolerable by all parties, and for most people, that number is merely whatever has been arrived at through a process of complex calculation, which could have been made simpler by just looking at a crystal ball. Perhaps the real motivation behind this stubborn belief in power is the protection that researchers wish to gain from assuming personal responsibility for undesirable business results. Indeed, "inadequate power" is perhaps one of the most frequently heard excuses for non-conclusive trials.

10.7 Statistical inference and testing for equivalence

The entire statistical decision-making theory of Neyman and Pearson is formal logic. Statistical inference, also known as statistical testing and test of hypothesis, is reasoning in formal logic. A hypothesis is a statement with exactitude. Statistical inference starts with a hypothesis, flowing out of which is an expectation quantified with a measure; then, the expectation is compared to the actual observations with intention to contradict the hypothesis.

To think in formal logic, the hypothesis must be exact. Before anything is known from the study, the hypothesis of no effect, known as the null hypothesis, is perhaps the only statement that could possibly be made with exactitude. For instance, the null hypothesis for comparing A and B is $A - B = 0$ or $A \div B = 1$, meaning that A and B are not different. Technically, the expression of difference or ratio is a matter of formality. If one has some idea how much A and B differ, the difference must be

precisely specified in order to test with formal logic. If the hypothesized difference is D, the hypothesis of $A - B = D$ may be transformed into a null hypothesis of either $A - (B + D) = 0$ or $(A - D) - B = 0$.

Inference in formal logic can only prove a hypothesis to be false. A null hypothesis is proven to be false upon detection of any difference. However, failure in detecting a difference does not logically prove equivalence. First, poor technique or inappropriate measure may be responsible for the failure. Had a better technique or an appropriate measure been used, the result might have been different. Second, failure in detecting a difference does not exclude other differences, and given the experimental nature of clinical research, there is no way to exhaustively test everything.

Statistical inference is too rigid to be useful in clinical studies. Not only is it extremely cumbersome to make exact hypotheses about everything under exploration, but also is it extremely restrictive that conclusion can only be drawn by proof of falsity. In the current statistics, all statistical tests use this formal logic and are designed to prove falsity of the null hypothesis by detecting a difference. This one-way logic favoring large differences presents a big problem for studies where the primary interest is to demonstrate equivalence, and the differences between treatment groups are expected to be small. For the same reason that the difference between treatment and active control tends to be smaller than that between treatment and placebo, statistical inference discourages the use of active controls. This lack of statistical tests for studies where by design the differences among treatment groups ought to be small is referred to as the problem of testing for equivalence.

There are no solutions to the problem of testing for equivalence within the domain of Neyman and Pearson. The widely adopted confidence interval approach is a pseudo-resolution. The key to the confidence interval approach is statistical power and a "clinically significant difference" declared before the trial. The technique itself is straightforward. In fact, when the exact difference between the comparing groups is given, a null hypothesis can be formed and a confidence interval is merely a different wording of usual test of hypothesis in terms of p-value. It is the argument used to justify the technique that requires scrutiny. To claim equivalence, the statistical test must have extraordinary

power to detect difference of arbitrary size. The reason is that if a powerful test fails to detect any difference, the difference, if any, may indeed not be significant. Then the question is how much difference is different. A recommendation of 20% has been widely adopted as a "clinically significant difference." If the difference between two means are less than 20%, and the power of the statistical test is satisfactory, the two treatment groups will be considered to be equivalent.

That pseudo-solution once again demonstrates the conflict between the exactness required in statistical inference and the uncertainty of scientific exploration. First, statistical power is a mirage and entirely non-observable. A power of 99% has no bearing on the actual outcome. A doctor joke perfectly makes the point. A sick man was admitted and telling his doctor that he had cancer. "Don't worry, my son, this time I guarantee you walk out of here alive." "How can you be so sure, Doc?" Knowing his poor prognosis, the man was very skeptical. "Well, the cure rate for your disease is 10%, and since nine of the ten patients I saw had died, you must be that lucky one." More troublesome is defining a "clinically significant difference." A clinical significant difference is what needs to be demonstrated in the study. It is absurd to declare a difference in research document and later defense it when the study proves otherwise. Moreover, a clinically significant difference is a very slippery concept and its real meaning can be rather obscured. Except for life and death, it is really disputable to define clinical significance with a single parameter.

The only solution to this problem is to abandon the rigid test of hypothesis formalism and to free the human reasoning power for true scientific inference. True scientific inference is presentation of facts and exercise of professional judgment. True scientific inference is highly flexible and its conclusions are not always black and white. In scientific inference, the formality of difference or equivalence is a matter of convenience, and the problem of testing for equivalence becomes a non-issue. Scientifically minded researchers are conscious of what effects are being examined, how the effects are measured, how strong the observed evidences are, and what the conclusion means for the benefit of human being. Such a complex human intellectual activity does not lend itself to be completely comprehended by a system of formal logic, nor to be fully represented by any numeric system and mathematics. There are times when a decision has to be made and a point of determination has to be

imposed. That is the time when the researchers or the representatives of different interests must take personal responsibility. Although statistical methods may be utilized to evaluate the strength of observed evidences effectively and to argue different opinions unambiguously, decisions can only be made by people, and in a democratic system, the majority should prevail, in principle.

10.8 The maximum likelihood technique

In my opinion, the maximum likelihood technique of Ronald A. Fisher is truly useful an analytic technique that statisticians may offer to clinical research. It is the core technical component underlying all the analytic methods presented in Chapters Four, Five, Six and Seven of this book. The technique utilizes mathematical distribution with undetermined parameters to represent the frequency distribution of data values. The distribution functions for all the data values are then multiplied together, giving rise to what is known as the likelihood function, and the parameters are determined by maximizing the likelihood function. The maximization guarantees that when the parameters so determined are substituted in the mathematical distribution functions, the resulting frequencies, on average, best represent the observed frequencies of the data values. A technical presentation of the maximum likelihood technique is given in Appendix B.

Both the maximum likelihood technique and the theory of Neyman and Pearson utilize mathematical distributions. In the theory of Neyman and Pearson, mathematical distribution is an entity that symbolizes the imaginary truth to determine errors. It is logically necessary that the data follow exactly the assumed mathematical distribution. There is no objective evidence to sustain the assumption of distribution, and therefore, the assumption is eternally vulnerable to criticism from authorities. In the maximum likelihood technique, the use of mathematical distribution is entirely technical. It does *not* require that the data follow any mathematical distribution, and verification of distribution assumption is totally unnecessary. It only requires that the mathematical distribution be admissible to all the data values and convenient so that the parameters are interpretable. For instance, if the interest is the mean and its standard deviation, the normal distribution with two functionally unrelated parameters is very convenient, with one representing the mean and the other the standard deviation of the mean. A mathematical distribution

without sufficient parameters for the purpose results in loss of information, much like scooping oceans with a teacup. An example is exponential and binomial distribution being used for the purpose of analysis of variance. These distributions have only one parameter, and when they are used to compute the mean and its standard error, there is no independent parameter to represent standard error once the mean is determined. This lack of independent parameters to represent two pieces of information results in an artificial dependence between the mean and its standard error, and the problem is referred to as overdispersion or underdispersion. On the other hand, a mathematical distribution with too many parameters for the purpose results in technical difficulties, much like drinking tea with a ten-gallon drum. The problem of ancillary parameters is due to the use of complex mathematical functions, and the so-called conditional distribution, marginal distribution, and partial likelihood are some of the struggles to simplify the mathematics.

10.9 Clinical research and statistical methods

Clinical research is exploration of the unknown world by humans. Measurement, as precise and objective as it can be, is only half of the researcher's duty. The more important half is judgment, which is ultimately up to the intuition and wisdom of human beings, no matter how fuzzy, subjective and emotional they may be. In exploration of the unknown world without reference points, consistency and operability are perhaps the only criteria to judge the results of clinical research. To establish consistency, a large database contributed and shared by the entire medical society is a necessity. The convention of paper publication of "statistically significant" studies should give way to standardization of study methodologies, sharing of data electronically, standardization of data visualization techniques, and exercise of quality control.

Statistics should only concern measurement for the purpose of clinical research and unambiguous communication. Statistical measurement should be as flexible as human reasoning and must not be a prisoner of human intelligence. A story told in my kindergarten actually makes the point. There was a man who measured his feet and went shopping for shoes. In the store he realized the measure was not with him. He ran back home and came back with the measure, but the store had closed. Folks asked why not try with his feet at the first place. The answer was "I trust

my measure better than my feet." The whole nonsense in the current clinical research practice discussed in this chapter is very much like that shoes-shopping man. It is entirely due to Neyman and Pearson who forced the highly flexible human intellectual activity into a rigid formal logic. The statistical theory of Neyman and Pearson was born lifeless, and it has been haunting us for about a century. We just have to have the intelligence and decency to sign the death certificate and move on to true scientific challenges in the twenty-first century.

11
Good Clinical Practice Guidelines

Summary

The ICH good clinical practice guidelines set an ethical and quality standard for clinical investigations submitted to the regulatory authorities in the European Union, Japan and the United States. Associated with GCP is an evolving body of technical guidelines with respect to clinical studies in general as well as in specific therapeutic areas. Most of these guidelines are published in Federal Register, and recent publications are available electronically through the US FDA. This chapter introduces the ICH good clinical practice guidelines concerning the role of physician as a clinical investigator and discusses the responsibilities of investigators, informed consent, protocol development and handling of adverse events.

11.1 A brief history

Misadventures with untested drugs and well publicized tragedies, as that of thalidomide, prompted the regulation of pharmaceutical products. A series of legislations in the United States in the middle decades of the twentieth century established the requirements for adequate and controlled clinical studies in development of pharmaceutical products in the US

jurisdiction. The US Food and Drug Administration (FDA) first developed a set of guidelines in 1970s, which defined the responsibilities of sponsors, investigators, monitors and institutional review boards. These guidelines, published within the Code of Federal Regulations and Federal Register, unfolded the worldwide development of good clinical practice (GCP) guidelines. In 1980s, GCP guidelines had been developed and published in the European countries, Japan, Canada and Australia. The emerge of world market and the shared concern of human ultimate interest drive the movement toward a global harmonization of the clinical study guidelines developed in individual countries. A milestone of this movement is the first International Conference on Harmonization (ICH) held in Japan in 1995.

The ICH guidelines for good clinical practice (ICH GCP) have been suggested for adoption to facilitate the acceptance of clinical data by the regulatory authorities in the European Union, Japan and the United States. In this chapter, we will discuss the ICH guidelines, particularly those concerning the role of physicians. Although these guidelines have been primarily developed to regulate the pharmaceutical industry whose focus is marketing pharmaceutical products, the principle of ethical collection of quality data applies to academic clinical research as well, where the primary interest is treatment options, not marketing pharmaceutical products.

The ICH GCP guidelines as well as other ICH technical guidelines were published in Federal Register. At the time of writing this book, those documents were available for the public at **http://www.fda.gov**, under Regulatory Guidance. The public is encouraged to comment on the guidelines in writing.

11.2 An overview

GCP concerns ethics. The purpose is to protect the right, safety, and well-being of trial subjects and to protect the public from the results of incredible trials. The guidelines set a standard to govern the design, conduct, monitoring, recording and reporting of clinical studies that are intended for submission to regulatory authorities. GCP guidelines themselves are not law, and compliance with GCP is voluntary in most countries. However, GCP guidelines are by and large initiated, sponsored

and eventually adopted by law enforcement agencies, like the US FDA. Therefore, for all practical purposes, they act as law. Violation of GCP may result in rejection of submitted data and denial of marketing. Discovery of fraud may lead to government sanction, and possibly criminal charges.

In essence, GCP guidelines are consisted of definitions of responsibilities to key players in clinical trials, standards of key operations, and outlines of critical documents. The key players are sponsors, investigators and institutional review boards (IRBs). Key operations are informed consent, selection of qualified personnel, data acquisition and flow, monitoring for safety and prompt reporting of adverse events, and auditing for quality assurance. Critical documents include protocol, informed consent form, investigator's brochure, and study report.

The working flow in a clinical trial starts with documentation of planned operations. Upon approval, the plan is implemented, and the implementation is documented to the degree that an auditor can reconstruct the entire operation from the documents. The documented implementation is then audited against the approved plan. The entire working flow is, therefore, documented plan, documented implementation, and documented validation. The logic behind this flow is this: First, you must know what you are supposed to do before you do anything. Second, you must do what you are supposed to do. Third, whatever you do must be auditable; in other words, you must be able to prove to others that you know and have indeed done what you are supposed to do. The idea of auditing prompts the vital role of documentation for the acceptance of clinical studies by regulatory authorities.

11.3 Benefits of compliance with GCP guidelines

GCP guidelines are highly operable. They are the backbone of standard operating procedures (SOP) in almost all international pharmaceutical corporations who aggressively pursue compliance. Compliance with the guidelines is no longer an option but a necessity for success in international commerce. Because studies in compliance with GCP are readily acceptable by authorities that have adopted GCP, compliance is the most efficient way to approach a broad market over

different jurisdictions. Quality products demonstrated by quality studies protect not only the public from ineffective and unsafe products, but also the pharmaceutical industry from poor public image and losses in lawsuits.

For clinical researchers, GCP guidelines must not be considered as government restrictions on industrial clinical trials. They are a rich source of information for everybody to design and conduct ethical and scientifically sound clinical studies. GCP guidelines help physicians to be knowledgeable to the processes and legal obligations in clinical studies, to write SOPs to sustain the consistency and efficiency of a research team in today's dynamic and diverse labor market, to gain credit and favorable publicity by presenting quality studies, and to avoid common mistakes and litigations. Compliance with GCP guidelines for clinical investigators is essential for physicians to stay in the mainstream of drug development business in the developed countries.

11.4 Investigators

An investigator is a person who actually conducts the trial. This person takes or supervises the care of trial subjects, and under the direction of this person, the trial product is administrated to trial subjects. An investigator is a key player in clinical trials. This player links the sponsor, IRB, trial subjects, site staff and monitor personnel. An investigator must be qualified by education, training and experience for taking medical care of trial subjects, must hold the responsibility to conduct the investigation in accordance to the approved plan, and needs to have adequate resources for fulfill the responsibilities.

11.4.1 Responsibilities

The investigator for a trial must know the investigational product by reading the investigator's brochure, protocol and any other product information from the sponsor. If the investigator delegates duties to other persons, she or he must be conscious of, preferably document, the duties delegated, and is responsible to inform the delegated persons of the trial product.

The investigator must prepare an informed consent form, preferably after consultation with the IRB who must eventually sign off the form. Usually, the sponsor provides a sample, which may be adopted or

modified. The consent form must disclose the experimental nature of the trial, procedures involved, treatments under comparison, the chance of being assigned to a particular treatment, expected benefits, known or foreseeable risks, and alternative managements other than the treatment under trial. The consent form must state that it is completely voluntary to participate in the trial, to continue with the trial and to withdraw from the trial, and whatever a subject chooses will not result in loss of right and benefit that the subject is otherwise entitled to. Payment and reimbursable expenses need to be prorated to ensure that participation in the trial is not unduly influenced by financial gain. The consent form must include statement on confidentiality and the protection of subjects' identification. In the meanwhile, the form must also request authorization, by signing the form, for legal access to the data for the trial or anything related to the trial. The tone of the writing should be neutral, and the content must be factual. The consent form must never sound like advertising from an automobile retailer.

The investigator is responsible to communicate with the IRB to obtain approval for the protocol, informed consent and advertising for patient recruitment. A protocol needs to be sent to the IRB for review and approval at least annually. The investigator must update the IRB with all newly developed safety issues, changes in trial conduction, deviation from the protocol for any reason and any other issues that may jeopardize the subjects' right and well being.

The investigator must uphold responsibilities by signing an agreement with the sponsor, and filing necessary forms to regulatory authorities, for instance, the FDA Form 1572. The investigator must agree to conduct the trial in accordance with the approved protocol. The protocol should not be deviated unless the trial subjects' interest is jeopardized and there is no instruction in the protocol to deal with the situation. Any deviations from the protocol must be reported to the IRB and sponsor in a timely fashion.

The investigator is responsible to obtain signed and dated informed consent from trial subjects. Sufficient time should be allowed for subjects to read and make a decision. Questions should be answered to the subject's satisfaction. The informed consent form must be signed and dated before any trial procedure is applied to the subject. This is a legal requirement.

The investigator is responsible to collect and report data as specified in case report forms (CRFs). The collection and recording of CRF required data must be accurate and complete. If the data on CRFs are derived from a source document, the data must be consistent with the source document. Any discrepancies must be resolved or explained. To maintain an audit trail, any correction or change must be documented, signed and dated, and the original entries must not be obscured. The investigator is responsible to retain all documents that relate to the trial activities at the site for at least two years after termination of the clinical development program from the sponsor.

The investigator is responsible to report adverse events to the sponsor. Serious adverse events must be reported immediately to the sponsor and IRB. Serious adverse events should be treated for the best interest of the subject. If that treatment presents a protocol violation, the sponsor and IRB should be notified. All correspondence must be put on writing. Attention must be paid to preserve blinding and not to violate subjects' confidentiality. It is preferred to use subjects' trial identification number in all correspondence.

The investigator is responsible to strictly follow the instructions from the sponsor on the storage, dispensing and disposition of trial product. An inventory of trial product must be maintained and kept in record. Collaboration with a pharmacy is often helpful in this regard. The accountability of trial product is vital for the credibility of the trial. Upon audit, the investigator must be able to demonstrate that the trial product was indeed dispensed in accordance to the protocol, and the trial product was only used on trial subjects.

11.4.2 Necessary resources for an investigator site

Investigators are usually identified by reviewing their education, training and experience in the medical specialty of the trial. However, even an expert with great enthusiasm in the trial may not qualify to be an investigator. Functionally, an investigator is not viewed just as a person but the representative of a team and resources available for the team. Although the investigator is held for responsibilities, daily activities are mostly carried out by a team of site personnel in a supportive facility.

The patient population under service is one of the most important resources. An adequate number of patients with the trial indication directly determine whether or not the quota of patient recruitment can be met within a limited time frame. As a candidate investigator, it would be most impressive if you can present statistics on the patients with the trial indication who are under your care or accessible by referral. Statistics in the local area is also helpful if you plan advertising to speed up patient recruitment.

To my best knowledge, clinical research is not taught in any school of any kind. Therefore, a demonstration of knowledge in GCP, training and experience in clinical research for the whole team will add a competitive edge. Delivery of quality work requires qualified work force. To maintain a consistent service over a long run with today's dynamic and diverse labor market, developing and maintaining SOPs, compliant with GCP and tailored for the facility, seem to make good business sense. Clinical research associates (CRAs) from any established pharmaceutical company who are experienced in site monitoring can be a rich source of information for laying down procedures and developing SOPs.

As a candidate investigator, you have to be aware of competing trials, trials that demand not only the time of your staff but the patients as well. A good tracking of workload, staff hours and cost is important not only for management of the team, but also for demonstrating staff availability and negotiating budget.

Information about IRB should be available. It includes the members, their qualification, services, availability and fee. Requirements of the IRB should be disclosed in case they present a problem for the trial.

A properly operated clinical laboratory should be available. Certification for operation, quality control measures, normal ranges, data flow and fee schedule should be carefully evaluated and presented if necessary. Special services, such as radiology, surgery, pathology, and hospital that are required for the trial or for the treatment of adverse events should be available. The arrangement with, and an assessment of, these services may have to be presented.

Reliable services from a quality pharmacy are invaluable to fulfill the investigator's responsibilities. A tour of the pharmacy, guided by a

competent pharmacist, should be planned at the initial site visit. The tour should show the facility for storage and dispensing of trial product and the security system. The tour must also show the system for tracking drug accountability, implementing randomization, and preserving blinding. Studies with dynamic randomization heavily rely on pharmacy services. Pharmacist's experience in clinical trial is highly appreciated. For trial with complex randomization and drug dispensing, the sponsor usually offers training to pharmacists. As an investigator, you should encourage or sponsor interested pharmacists to take those paid learning opportunities.

11.5 Institutional review boards

An institutional review board (IRB) or an independent ethics committee (IEC) is a group of people, usually designated by or affiliated with an medical institution, who review and approve proposed and ongoing clinical trials. An IRB or IEC should have at least five members, of whom one must be independent of the institution or trial site, one must be of nonscientific profession, for instance, an attorney, and one must be of scientific profession. An IRB or IEC must have both men and women. The objective of IRB is to protect the rights, safety and well being of trial subjects and to provide public assurance of that protection.

The IRB or IEC for a trial is responsible to assess the risks to trial subjects on a continuous basis and make decisions on whether the risks to trial subjects are minimized. Protocol, investigator's brochure, available safety information, progress reports, and any updates or amendments of the above documents must be reviewed. Following the assessment, the IRB must offer its opinion and document its decision on approval, modification, disapproval, suspension or termination of the proposed or ongoing trial.

The IRB or IEC is responsible to ensure that trial subjects are fully informed. The informed consent form must be reviewed against legal requirements, government regulations, GCP guidelines and all other ethical concerns. Special attention must be paid to vulnerable subjects, payments to subjects, and advertising for subject recruitment. Vulnerable subjects are children, the mentally or physically impaired who might not be able to fully understand the implications of being in the trial and make voluntary decisions for their best interest. The IRB must identify

vulnerable subjects and ensure that ethical issues are addressed and their interest is protected. The IRB must curb excessive payments to subjects. Excessive payments may invalidate informed consent and be considered as coercion. Regional or national statistics on payments to trial subjects may be a useful reference.

The IRB or IEC communicate with the investigator, not the sponsor in general. Their operation is to request information, review the information, make a decision and inform the investigator of the decision. IRB operations are regulated and subject to inspection from regulatory authorities. Thus, the members must be familiar with all legal requirements, government regulations, and GCP guidelines, and operate accordingly. Every operation must be documented and available for audit and inspection. The documents should be retained for at least three years after termination of the clinical development from the sponsor. If the IRB is found guilty of noncompliance, the whole institution may be sanctioned by regulatory authorities from clinical trials.

11.6 Sponsors

A sponsor is an individual or any business entity who initiates, manages, finances and potentially benefits from a clinical trial. There is a long list of sponsor's responsibilities in ICH GCP guidelines. The main focus here is development of protocol and handling of adverse events. These are the butter and bread of sponsor physicians in daily clinical trial business.

11.6.1 Development of protocol

Protocol is a comprehensive document that describes a trial from the idea to the very details of operation. The content generally includes objective, design, patient population, parameters to be observed, and procedures for data acquisition and analysis. The information in the protocol will be used by regulatory authorities to inspect for compliance, by IRBs to assess ethics, by investigators, as part of the contract with the sponsor, to carry out the trial in the set standard, by CRAs to design CRFs and monitoring guidelines, by data management team to design database, and by statisticians to initiate an analysis and reporting plan.

The ICH GCP guidelines recommend the contents of protocol. The following is a commonly used format to organize those contents:

Title

- Describe the drug, objective, design, indication and population.

Contacts

- Names and titles, numbers and addresses. Emergency contacts must be included.

Summary (Optional)

- Briefly describe the treatment, design, time frame, population, and endpoints.

Background

- Summary of the current findings, clinical and non-clinical, with attention to safety and potential benefits.
- Highlight the rationale of this trial in the above context.

Objective

- Specify what the trial is designed to demonstrate and claims to be made.

Trial Design

- Draw a flowchart. Describe in detail:
  - the phases, such as run-in, screening, randomization, follow-up and trial stop,
  - the schedule of each phase,
  - data to be collected in each phase,
  - the schedule of procedures and data acquisition.
- Trial subjects:
  - Specify the number of subjects in each treatment group.
  - Specify the number of centers initiated and a rough quota to each center if it is a multicenter trial.
  - Delineate the inclusion and exclusion criteria by a series of tick boxes.
- Treatment:
  - Give complete prescription information.
  - Specify permitted and prohibited concomitant treatments.
  - Measure compliance with, for example, diary cards, returned drugs.
  - Prompt documentation of drug accountability.

Efficacy

- Define every efficacy parameter and its measurement.
- Discuss, optional but helpful, how the efficacy measures translate to patient benefit.

Safety

- Define safety parameters, including critical laboratory values.
- Define serious adverse events and guidelines for assessing causality.
- Define criteria for reportable adverse events and the time frame.
- Provide guidelines for the contents and extent of safety data.
- Describe the system for the flow of safety data.

Withdrawal

- Define criteria for patient withdrawal.
- Provide guidelines for investigating withdrawals with attention to the time and relation to lack of efficacy or intolerable adverse reactions.
- Specify a follow-up plan to withdrawn subjects, if necessary.

Data management

- Describe the SOPs for handling of CRFs, data entry, quality control (QC) and query.
- Define data flow and system validation.

Statistics

- Specify sample size, design and the randomization process.
- Define data for analysis, such as the intent-to-treat, evaluable and per-protocol populations.
- Define primary and secondary parameters to avoid unnecessary troubles.
- Make a statistical analysis plan and declare a significant level.

Special sections

- Pharmacokinetic studies, pharmacoeconomic studies.

Quality control

- The SOPs for trial monitoring and internal auditing.

Technical issues

- (Appendices).
- Instruction for shipping and handling of trial products.

- Description of test procedures, specimen handling and labeling, etc.

In the pharmaceutical industry, the development of trial protocols is directed by an overall product development plan with focus on marketing. A protocol is an assembled document from contributions of a team, consisted of physicians, CRAs, CRF design group, data management, and statisticians. A medical writer usually coordinates the team activities and physically assembles the documents from the involved parties. The draft protocol normally circulates for several times for reviews and revisions. The final draft is presented to the management team for sign off.

11.6.2 Handling of adverse events

Adverse event or adverse experience is any untoward signs, symptoms, or disease that is temporally associated with, not necessarily caused by, the use of a medicinal product. The GCP guidelines require expedited reporting of serious and unexpected adverse drug reactions to regulatory authorities within a time frame, prompt notification of the discovery to the investigators and IRBs, and amendment of the investigator's brochure with the new information.

The sponsor physician is responsible to identify adverse events that require expedited reporting. A reportable adverse event must be serious, unexpected and possibly caused by the investigational product. An adverse event is defined to be serious if it results in death, is life threatening, requires or prolongs hospitalization, results in persistent or significant disability, or is a congenital defect. Seriousness is defined by the outcome. It must not be confused with severity, which describes the intensity. An adverse event is defined to be unexpected if the nature or severity is not consistent with the current knowledge in relevant documents. For investigational products, the information in investigator's brochure is commonly used to determine unexpectedness. If a compound is investigated in multiple formulations or on multiple indications, it is important that the investigator's brochure is specific for the formulation or indication under trial. Unexpectedness is product-specific and indication-specific. To qualify for expedited reporting, a causal relationship between the adverse event and the investigational product must be assessed to be a

reasonable possibility. The sponsor physician must consult with the reporting investigator in assessment of the adverse event. In general, the sponsor physician should not overrule the investigator's assessment. If the treatment is blinded, the assessment of causal relationship requires breaking the blind. The extent of unblinding should be minimized to avoid compromising the validity of the trial.

If an adverse event is fatal or life threatening, the sponsor must report it, by all means, to regulatory authorities as soon as possible but no later than 7 days after first knowledge that the case qualifies for expedited reporting. The initial report should be followed by as complete a report as possible within 8 additional days. If an adverse event is qualified for expedited reporting but not fatal or life-threatening, the sponsor must report it as soon as possible but no later than 15 days after knowledge by the sponsor. The initial report must contain the information for identifying the patient, the source of report and the investigational product. It must also contain the information that describes the event and qualifies it to be reportable. To ensure compliance, the timing of the reporting process must be documented.

GCP guidelines list the following key data elements to be included in expedited reports of adverse events. They read as follows:

1. **Patient details:**
    - Initials,
    - Other relevant identifier (clinical investigation number, for example),
    - Gender, age or date of birth, weight and height.
2. **Suspected medicinal product(s):**
    - Brand name as reported, International Nonproprietary Name (INN),
    - Batch number,
    - Indication(s) for which it was prescribed or tested,
    - Dosage form and strength,
    - Daily dose and regimen (specify units: mg, mL, mg/kg, etc.),
    - Route of administration,
    - Starting date and time, stopping date and time, duration of treatment.

3. **Other treatment(s):**
    - For concomitant medicinal products and nonmedicinal product therapies, provide the same information as for the suspected product.
4. **Details of suspected adverse event:**
    - Full description of reaction(s) including body site and severity, as well as the criterion for regarding the report as serious, if possible, specific diagnosis.
    - Start date and time, stop date and time, and duration.
    - Dechallenge and rechallenge information.
    - Setting, e.g., hospital, outpatient clinic, home, nursing home.
    - Outcome: Information on recovery and any sequelae; specific tests and/or treatments and their results; for a fatal outcome, cause of death and a comment on its possible relationship to the suspected reaction; autopsy or other post-mortem findings if available; any other information relevant for assessing the case.
5. **Details on reporter of event:**
    - Name, address, telephone number and profession (specialty).
6. **Administrative and sponsor details:**
    - Source of report: spontaneous, from a clinical investigation, the literature.
    - Date of the event report received, country where the event occurred.
    - Type of report filed to regulatory authorities (initial, follow-up, etc.).
    - Name, address and contact information.
    - Regulatory code or number for marketing authorization dossier or clinical investigation (IND and NDA numbers).

12
Data Management in Clinical Research

Summary

A focus of this chapter is to explain the qualitative nature of clinically useful information for management of individual patients. It is advocated that integration of information is more important than the mere precision of measurement. Under this thinking is the promotion of clinically meaningful scales and data representing judgment. The result is improvement of efficiency in data acquisition, data flow, and data storage without compromise of information. Another focus of this chapter is to introduce a few basic ideas of data management to individual physicians who cannot rely on professional services for data management. Those tips may help them organize their data better and translate their hard work more readily into analyzable data.

12.1 Perspective

Clinical data management (CDM) is a profession. The book edited by Rondel *et. al.* is a good introduction to that profession. My perspective in this chapter is to explore the collection and management of clinically useful data. This initiative originates from my experience in the analysis

and reporting of clinical trial data in the pharmaceutical industry and with individual physicians. Clinical studies generate high volume of data, and clinical data management is very labor-intensive and time-consuming a business. I saw boxes after boxes of paper report docked for review and submission. I saw working bees, female mostly, in rows after rows of cubicles in data management department, busy between computers and piles of CRFs. I shared the frustrations of data cleaning and querying. However, my worst experience was working with individual physicians who conducted clinical research on a shoestring budget and had no slightest idea on data management.

Ignoring data management in clinical research is like an all-hated parsimonious man who bargains hard for a diamond and then throws it in a trash bin. On the other hand, unthoughtful collection of everything is like mixing diamond and trash in a security box in the Federal Reserve. The goal of this chapter is going through a thinking process with focus on defining clinically useful information and providing rudimental guidelines for individual physicians. The purpose is to avoid waste of resource on generating large amount of data of little use and translate the hard work of individual physicians more readily into analyzable data.

12.2 Grading with clinically meaningful scales

There has been a constant struggle to quantify subjective responses in clinical trials. Grading has been a widely accepted practice. However, the scale of grading remains controversial. The goal of this section is to promote grading subjective response in clinical meaningful scales and to argue against the use of equivocal scales in grading.

12.2.1 Pain scores and visual analog scale

A ridiculous practice in history taking is to ask the patient to grade pain from 0 for no pain and 10 for the worst pain. Even worse is the adoption of visual analog scale (VAS). A visual analog scale is usually a straight line with one end representing the minimum and the other the maximum. The patients are instructed to strike a mark on the line to indicate the severity of their symptoms, and a distance is used to quantify to subjective responses. By visual analog scale, the severity of pain, for instance, is measured with a numeric number, say, 13.5cm.

It is seriously doubted that grading with scales like pain score from 0 to 10 or visual analog scale help us collect any clinically meaningful information from the patients. The reason is that those scales have no specific meaning, and people do not think in terms of numeric numbers without clear meanings attached to them. At least it is my experience that most patients are stunned when asked to grade their symptoms from 0 to 10, and they hesitate to give an answer simply because they do not know what those numbers mean to them. If the patient gives an answer right away, you immediately know that he or she is a professional patient, and you should watch out for drug seeking. Indeed, what does a pain score of 5 mean? How much more pain does score 8 represent than does score 5? Should we intervene if the pain score is 9.5? Without clear answers to these questions, the pain score is totally useless for clinical practice.

Not only are the data so generated carry no clinically useful information, visual analog scale or scales from 0 to 10 with no clearly defined meaning to each number greatly increase the volume and complexity of data operation. Imagine that the site personnel have to measure the distance and enter the number onto the CRF. If the site person inaccurately measures the distance or enters a wrong number but within the acceptable range on the CRF, this error may slip through computer check and is only detectable upon audit. Furthermore, even if the site person has done anything right, the decimal point, for instance, may be blurred during the transportation of paper CRFs or completely missed by the data entry technician. Finally, floaters take more computer space than integers, although nowadays this is a relatively trivial issue.

If we carefully review the history, it is not difficult to find out that the real motivation behind this strange practice has to do with statistical testing. Because of the mathematical complexity associated with discrete numbers, there is a lack of well-developed statistical tests by the theory of Neyman and Pearson for discrete numbers. The asymptotic approximation is generally required to construct a sensible test, and for some, this mathematical approximation is not acceptable. Some argue that abundant discrete scales or visual analog scale can pick up subtle differences even though a particular point has no specific meaning. Indeed, by using continuous data, there might be a slightly better chance to obtain a small p-value from statistic test for claiming significance, for instance, between the mean pain scores of 14 and 15. The point is that a continuous or near-

continuous scale gives no clinically useful information, and because people do not think in terms of numbers for their symptoms, whatever numbers the patients are forced to pick may not well represent the true messages they intend to convey.

12.2.2 Clinically meaningful scales

In my opinion, clinically useful scales should closely associate with clinical assessment and intervention. Instead of from 0 to 10, for instance, one may simply grade the severity of nausea with

- 0 for no nausea,
- 1 for nausea but oral intake is adequate and no malnutrition,
- 2 for nausea and nutrition support is required.

As another example, pain, depending on the clinical scenario, may be simply graded with

- 0 for no pain,
- 1 for pain but tolerable, knowing the physiology of pain and the adverse effects of analgesics,
- 2 for pain that requires relief with analgesics despite the knowledge of pain physiology and the adverse effects of analgesics,
- 3 for pain but drug craving is suspected or documented.

The percentage of patients in each category can be directly used by clinicians to calculate the benefit and risk when they recommend the treatment to individual patients. A typical calculation is this. The durable response rate with this best available treatment is 10% based on reliable studies. The risk of life-threatening adverse events associated with this treatment is 50%. However, the chance of survival within a year without treatment is 5%. Therefore, the balance is a 5% gain in survivorship if treat but a 50% risk of suffering serious illness from the treatment. A statement, like 0.5 improvement in mean score, does not help in this kind of calculation. In addition, because scales based on clinical assessment and intervention are clearly defined and meaningful for both patients and physicians, the result of study is much more reproducible, and the data can be audited by referring to the assessment and intervention in the source document.

From the perspective of data management, a clearly defined scale system saves tremendous manpower. The scales can be easily illustrated with few tick boxes on the CRF, which will simplify data entry and reduce

the chance for errors. In fact, with electronic CRFs, the site personnel may actually select the appropriate options while interviewing the patient, and then print out the relevant pages to be included in the source documents after reviewed and signed by the physician. If the initial entry is correct, this virtually rules out any chance for error in the entire data flow.

12.3 Raw data and data representing judgment

Patient diaries, routine laboratory values, records of continuous monitoring, and machine generated reports often bring in tremendous volume of data, most of which are of little use for producing clinically relevant information. This is the time when we have to decide what to keep in the database, for instance, every byte of the electronic signals that produce an EKG versus clinically relevant diagnosis from an independent cardiologist who reads the EKG. While electronic signals are hard data, the diagnosis from an EKG reader is a judgment call.

My view is that data representing judgment calls are at least as good as raw data from routine procedures and monitoring. My argument for this view is that the art of medicine is more about integration of information than the mere precision of measurement. With so much unknown influences and uncertainty, sticking with ballpark numbers is not as earth-shuttering a disaster as missing critical information and losing sight of the global picture. A qualified physician generally has a better chance to interpret the data from the patient she or he is actually caring than does an in-house officer with a fancy computer. Although human brains are losing the match in the speed of calculation to computers, human brains still do much a better job in integration of information than computers do. When comes to clinical evaluation of patients, that cliché still holds true that system performance depends more on system optimization than the optimization of individual components.

Therefore, instead of everything on diary cards, we may keep the diary cards as source documents and only enter the results of diary review at each visit by the site investigators. The result may be categorized into, for instance, compliance with treatment, noncompliance, stable disease, unstable disease, and etc. Routine laboratory should be done locally and reviewed by the site investigator in reference to the local norms. The result may be recorded as being normal without further values and

abnormal with values grouped into panels. For machine-generated data as those from EKG, only critical parameters with clear clinical interpretations should be selected and kept in the database. If such parameters are not readily available, diagnoses in standard terminology, preferably from a standard dictionary, may be kept in the database. Subjective evaluation is most reproducible if the criteria are clear cut, in conformance with general practice, and formatted into a series of simple questions with answers restricted to yes or no.

12.4 Data management for physicians on a shoestring budget

For individuals, data management is not so much about the data. It is about ideas and structures. The purpose of this section is to discuss some basic ideas and structures for individual physicians who would like to accumulate data efficiently and have their data ready for analysis.

Unlike stethoscope, which is worn more as a symbol of medical profession than a tool for diagnosis, computer is indispensable for management of information. While fast and expensive computers are good for telecommunication and fascinating games, a personal computer with a good monitor and supporting video card is usually more than enough for managing clinical data for individual physicians. Tools for data management constantly change, but the idea and data structure seldom do. Therefore, the focus here is to discuss some basic ideas of data management and data structure. For all practical purposes, a spreadsheet is sufficient. I chose the Microsoft Excel© because it is widely available and supports dynamic data exchange with other database and statistical analysis software packages. If you never used a computer, you may start now by learning few basic operations: open a file, edit the data, save the file, and copy the file to a floppy disk as backup. Two operations are essential to data editing: enter and delete. Do not even try to read manual, consult your high school children instead. There is no need to chase the fad of computer market, and do not be lured by unbelievable demonstrations.

The key in constructing databases is unique identification of information. Suppose you do not have an idea exactly what data you are going to collect. An open structure may be the only choice. The structure is based on the assumption that

- the data are measurements from individual patients,
- each patient may be assessed over a period of time, and
- at each time the same measurement may be repeated several times.

Each piece of information is uniquely identified with patient, the time of patient evaluation and data acquisition, the sequence of repeated measurement, and a description of the measurement. Because patient is the primary identifier of patient related information, it is convenient to construct a separate database for patient identification:

Patient Identification File						
ID	**Name**	**SSN**	**MRN**	**Sex**	**Race**	**DOB**
1	Fritz	123-45-6789	123456	M	A	12/23/2005
2	Thee	234-56-7891	234561	F	B	09/15/1891
3	Caatt	345-67-8912	345612	F	W	04/21/1966
SSN: social security number, MRN: medical record number						

Whenever a new patient is included in the study, you just simply update the database with that patient's identification information. In this file, each patient is uniquely identified with the patient's ID number, and this number can used in or linked to other data files. The following is a database in open structure:

A Database in an Open Structure				
ID	**Date**	**Series**	**Field**	**Value**
1	12/15/1998	1	PFT	Abnormal
1	12/15/1998	2	PFT	Normal
1	09/25/1999	1	CT cap w/c	POD
1	10/30/2010	1	CT cap w/c	Stable
2	10/12/1978	1	BAL	Lymphocytes
2	12/15/1988	1	ABG	7.44/35/65/90%/7
2	12/15/1988	2	ABG	7.43/45/76/97%/9
3	10/30/2010	1	CT h wo/c	B.G. hemorrhage

Each piece of information is uniquely identified with a combination of ID, Date, Series, and Field.

But this database is too crowded. It is going to be your nightmare at the time of analysis when you have to gather common things together. If you have some idea on the information you are going to collect and accumulate, you may break down the information into panels. This is the

idea of fragmentation. For instance, you may isolate AGBs and group the parameters into another database. With spreadsheets, you simply open a new sheet and set up the database. If you are still not sure whether or not you may add other parameters later on, you may still use an open structure:

ABG Database with an Open Structure				
ID	**Date**	**Series**	**Field**	**Value**
1	12/15/1998	1	Ph	7.44
1	12/15/1998	1	PCO_2	35
1	12/15/1998	1	PO_2	65
1	12/15/1998	1	O_2 sat.	90%
1	12/15/1998	1	Hb g/dL	7
1	12/15/1998	1	FiO_2	21%
1	12/15/1998	2	Ph	7.43
1	12/15/1998	2	PCO_2	45

However, if you have very much determined the parameters for a database, you may adopt a more closely defined database structure:

ABG Database with a Close Structure								
ID	Date	Series	Ph	PCO_2	PO_2	O_2 sat.	Hb g/dL	FiO_2
1	12/15/1998	1	7.44	35	65	90%	7	21
1	12/15/1998	2	7.43	45	76	97%	9	21

In summary, the key for building a database is unique identification of information, and when it is feasible, fragmentation of information into panels. The panels are linked with a combination of identifiers. Whether you choose an open or close structure is a matter of convenience. If you are not sure of the parameters, an open structure allows you to add whatever parameters to whichever patient. A close structure is more convenient if you know the parameters and chance of adding more is small.

Few tools may make data management easier. You might want to build or purchase few dictionaries, preferably electronic, for diagnostic codes, drug codes, and perhaps adverse event codes. Numeric data are ready for analysis. It is non-standardized text entries that are analyst's nightmares. For instance, today you enter CAD under in the field of

DIAGNOSIS in your database, few weeks later, you may enter coronary a. disease, and years later, you may enter ischemic heart disease. Although the entries are meant to represent the same disease entity, a computerized analysis will present these as separate entities. A dictionary will help you standardize your text entries and make your data ready for statistical analysis with computers.

If an electronic dictionary is used, I suggest that you enter the code in your database. If the dictionary is on paper, you need to enter both code and text. Commercial dictionaries are constantly updated. The text matched to a code today may not be the same tomorrow. By including code in your database, you can always match the code to whatever text that is current. On the other hand, if you only enter the text in your database, you have to manually change it if the updated dictionary adopts a new nomenclature system. With code in the database, sophisticated programs can actually link the code automatically in the background to the text in the dictionary so that the meaning of the code in your database is automatically updated when the dictionary is updated.

The last thing is safety. It is a good practice to backup your electronic files from the hard drive to numerous floppy diskettes and keep two in your office, two in your briefcase, one in your car, and one at your home. This reduces the chance of total loss of data. However, when the master files are updated, all the backups need to be refreshed, and if some of them are left unchanged, shuffling diskettes may generate a good deal of confusion. If network drives are available, it is safer to store data on network drives than on local hard drives. Network drives are usually maintained by professional services with adequate safety measures. However, before you store your data on a professionally maintained network device, you need to first work out a confidentiality agreement with the service provider.

12.5 Global clinical data management

With the advance of telecommunication technique, it is probably about the time to entertain the idea of global clinical data management with flexible services to individual clinical researchers. It is probably too ambitious to implement a centralized system managed by a single business entity. To me, it is more practical and flexible to implement a global

standard to individual researchers for the storage, quality control and sharing of data from clinical trials. The result would be a web of databases and a central registration system for clinical studies, saving the need for searching engines. With such a web, future medical publication will not be just few pages of printed material, but an invitation to the data. I believe that only the United States government has the capacity to initiate a project like this by requiring mandatory adoption of a standard for government funded research projects and trials for marketing government controlled products to the public. The data with assured quality and more efficient use of scientific information should justify the initial cost over a long run.

Appendices

A Get results with SAS®

SAS® is a computer software package. This appendix shows how to use SAS to carry out the analyses presented in Chapters Four, Five, Six and Seven.

CENTER	PID	VISIT	TRT	BSL	FEV1
0001	001	1	D	5.35	5.32
0001	001	2	D	5.35	6.25
0001	001	3	D	5.35	4.98
0001	001	4	D	5.35	5.72
0001	001	5	D	5.35	6.84
0001	002	1	P	4.78	5.30
0001	002	2	P	4.78	5.92
0001	002	3	P	4.78	5.34
0001	002	4	P	4.78	6.02
0001	002	5	P	4.78	5.86
0001	003	1	P	5.67	4.36
0001	003	2	P	5.67	4.83
0001	003	3	P	5.67	5.52
0001	004	1	D	4.96	6.82
0001	004	3	D	4.96	6.92
0001	004	4	D	4.96	5.91

A.1 SAS-ready data

Data organized in the list format are ready for SAS. The above table shows the list format for data from a typical multicenter study, where patients in each center are randomly assigned to treatment (TRT) groups

® SAS is a registered trademark of SAS Institute, Inc.

and then followed at a sequence of visits. Each record (row) is uniquely identified by the combination of CENTER, PID (patient id) and VISIT. TRT (treatment) and BSL (baseline) are patient-specific; FEV1 is visit specific. Columns are referred by names on the top.

A.2 PROC GLM for the analysis of variance

For all practical purposes, PROC GLM is the procedure of choice for the analysis of variance. GLM stands for general linear models. PROC GLM can be used to compute the means and their standard errors, the least squares means and their standard errors, and the mean sums of squares for building an ANOVA table.

Suppose the data illustrated in section A.1 are ready for analysis. For the analysis specified in the linear model,

$$\text{fev1} = \text{baseline} + \text{center} + \text{treatment} + \text{center-treatment interaction} + \text{residual},$$

the following SAS code produces the desired summary measures at each visit:

Means, Least Squares Means, Standard Errors and ANOVA Table

```
PROC GLM OUTSTAT=ANOVATAB DATA=in_a_1;
  BY VISIT;
  CLASS center trt;
  MODEL fev1 = bsl center trt center*trt / SS4;
  LSMEANS trt / stderr pdiff out=LSMEANS;
  OUTPUT OUT=MEANSTDR PREDICTED=mean STDP=stderr;
RUN;
```

The means and their standard errors are listed in the output data file MEANSTDR. The least squares means of treatment groups and their standard errors are output to data file LSMEANS. The mean sums of squares for building the ANOVA table are output to data file ANOVATAB. For the analysis specified in this model,

$$\text{fev1} = \text{baseline} + \text{center} + \text{poly(visit,2)} + \text{treatment} + \text{treatment-poly(visit,2) interaction} + \text{residual},$$

the mean response profiles over the time course are represented with quadratic curves. The following SAS code produces the desired mean responses and their standard errors for all unique combinations of the independent variable values.

Means, Standard Errors and ANOVA Table

```
DATA POLY; SET A1;
  v1=visit;
  v2=visit*visit;
PROC GLM OUTSTAT=ANOVATAB DATA=poly;
  CLASS center trt;
  MODEL fev1 = bsl center v1 v2 trt trt*v1 trt*v2 / ss4;
  OUTPUT OUT=MEANSTDR PREDICTED=mean STDP=stderr;
RUN;
```

The mean response profiles are best visualized with graphical display of the means and their standard errors from the output data set MEANSTDER.

PROC GLM can also be used to compute the three types of measures discussed in Chapter Four, section 4.3. To get type I measures for the effects of center, treatment and center-treatment interaction, each factor is put in the model at a time:

Type I Measures: Means, Standard Errors and ANOVA Table

```
PROC GLM data=a1;
  CLASS center trt;
  MODEL fev1 = center / ss4;
  LSMEANS center / stderr pdiff out=LSMEANS;
PROC GLM data=a1;
  CLASS center trt;
  MODEL fev1 = trt / ss4;
  LSMEANS trt / stderr pdiff out=LSMEANS;
PROC GLM data=a1;
  CLASS center trt;
  MODEL fev1 = center*trt / ss4;
  LSMEANS center*trt / stderr pdiff out=LSMEANS;
RUN;
```

To get type II measures for the effects of center and treatment, the two factors are put in the model together without interaction effects:

Type II Measures: Means, Standard Errors and ANOVA Table

```
PROC GLM data=a1;
  CLASS center trt;
  MODEL fev1 = center trt / ss4;
  LSMEANS center trt / stderr pdiff out=LSMEANS;
RUN;
```

To get type III measures for the joint effects of center and treatment, the two factors and their interaction are put in the model:

Type III Measures: Means, Standard Errors and ANOVA Table

```
PROC GLM data=a1;
  CLASS center trt;
  MODEL fev1 = center trt center*trt / ss4;
  LSMEANS center*trt / stderr pdiff out=LSMEANS;
RUN;
```

A.3 The four types of sum of squares in PROC GLM

PROC GLM allows users to choose from four types of sum of squares to build an ANOVA table. Different from the three types of measures discussed in Chapter Four, section 4.3, these four types of sum of squares often confuse novice users. They are better explained with an example. Suppose that we are interested in the effects of center, treatment and center-treatment interaction. Those four types of sum of squares may be viewed as four different ways to attribute data variations to the factors under analysis. The type I sums of squares are the additional variations attributed to a factor after taking into account the effects of other factors before it. When the effects of multiple factors are analyzed simultaneously, their type I sums of squares depend upon the order in which these factors are entered in the model statement of PROC GLM. For the analysis specified in the following linear model,

responses = center + treatment + center-treatment interaction + residuals,

the following table summarizes the interpretation of the three effects as measured by their type I sums of squares:

Center effects:	Variation due to center only
Treatment effects:	Additional variation due to treatment after center effects
Center-treatment interaction effects:	Additional variation due to the interaction effects after center and treatment effects

Because type I sums of squares are order-dependent, interaction effect must not stay ahead of their single components. For instance, if the order of the factors in the model is changed to

responses = center-treatment interaction + center + treatment + residuals,

the type I sums of squares of center and treatment effects will be all zero. The following table explains the reason:

Center-treatment interaction effects:	Variation due to the joint effects of center and treatment
Center effects:	No variation to center after the joint effects
Treatment effects:	No variation due to treatment after the joint effects

The type II sums of squares are additional variations after taking into account the effects of all other factors except for interaction effects:

Center effects:	Additional variation due to center after treatment effects
Treatment effects:	Additional variation due to treatment after center effects
Center-treatment interaction effects:	Additional variation due to the interaction effects after center and treatment effects

Type II sums of squares have nothing to do with the order in which the effects are entered in the model statement. The types III and IV sums of squares are additional variations after taking into account any other effects, including interaction effects:

Center effects:	additional variation due to center after treatment and center-treatment interaction effects
Treatment effects:	additional variation due to treatment after center and center-treatment interaction effects
Center-treatment interaction effects:	additional variation due to the interaction effects after center and treatment effects

Type IV may be preferred to type III sums of squares when the analysis includes interaction effects and there are missing observations in certain treatment arms at certain centers. Suppose the following table represents the distribution of patients in a study:

Treatment	A	B	C
Center 1	100	5	0
Center 2	5	100	50
Center 3	75	35	65

There is no observation in treatment group C at center 1, and that cell is known as an empty cell. In the presence of empty cells, the type IV sums of squares are computed from non-empty cells that both define the effects of interaction and make balanced contributions to main effects. The type IV sum of squares for the effects of treatment, for instance, may be computed from the data in centers 2 and 3, where the effects of center-treatment interaction can be well defined and each cell makes fair contribution for the estimate of treatment effects. However, type IV sums of squares are not unique. Users need to carefully examine the estimable functions associated with each type of sums of squares, which may be requested with the E1 through to E4 options of the model statement.

The four types of sums of squares can be easily obtained with the SS option:

Types I, II, III and IV Sums of Squares

```
PROC GLM;
  CLASS center trt;
  MODEL response = center trt center*trt / ss1 ss2
ss3 ss4;
RUN;
```

A note of caution is that when several types of sums of squares are specified for the same effects, PROC GLM compares all types of sums of squares to the same residual sum of square. This may not be always desirable for the type II sums of squares when there are interaction effects in the model.

Without interaction effects, all four types of sums of squares are the same. Types III and IV sums of squares are the same if there are no empty cells. When there are empty cells, the user needs to carefully examine the estimable functions. The balancing property of type IV sums of squares may be preferred.

A.4 PROC MIXED for mean and individual profiles

PROC MIXED allows for a matrix, as opposed to a scalar in PROC GLM, to represent the residual mean sum of squares, which has been advocated by some statistical authorities to be necessary for longitudinal data. For comparing the effects of treatment at each visit, the following analysis,

$$\text{fev1} = \text{baseline} + \text{center} + \text{visit} + \text{treatment} + \text{visit-treatment interaction} + \text{visit-treatment-center interaction} + \text{residual},$$

can be carried out with the following SAS code:

Means, Standard Errors and ANOVA Table

```
PROC MIXED NOBOUND METHOD=ML EMPIRICAL;
  CLASS pid visit center trt;
  MODEL fev1 = bsl center trt visit visit*trt
               visit*trt*center / PM;
  REPEATED / SUBJECT=pid TYPE=UN;
  LSMEANS visit*trt / DIFF;
  MAKE 'TESTS' OUT=tests;
  MAKE 'LSMEANS' OUT=lsmeans;
  MAKE 'PREDMEANS' OUT=predmean;
RUN;
```

The REPEATED statement defines the residual matrix. Although an unstructured residual matrix, specified with UN, is advocated to be more "natural", it is fairly arbitrary to choose whatever a matrix to expedite

computation, as long as the EMPIRICAL option is specified in the PROC statement. The EMPIRICAL option directs the program to compute the standard errors directly with the residuals. The standard errors computed as such are robust to a variety of matrix structures that user may choose to represent the residual mean sum of squares. The NOBOUND option, a pure technical issue, directs the program to accept negative variance components to preserve the relationship among the components. The MAKE statements create three data sets: TESTS for building an ANOVA table, LSMEANS for the least squares means and their standard errors, and PREDMEAN for the means and their standard errors.

PROC MIXED can be used to compute individual response profiles. This is achieved by defining appropriate random effects in the analysis. In the analysis of variance, fixed effects define means whereas random effects define individual effects and they are measured with the deviations of individual effects from the means. The following model,

fev1 = baseline + center + poly(visit,2) + treatment + poly(visit,2)-treatment interaction + poly(visit,2)-treatment-center interaction + RANDOM{intercept + poly(visit, 2)} + residual,

Average and Individual Response Profiles

```
DATA INDPROF; SET A1;
  V1=VISIT;
  V2=VISIT*VISIT;
PROC MIXED NOBOUND METHOD=ML EMPIRICAL;
  CLASS pid center trt;
  MODEL fev1 = bsl center trt v1 v2 v1*trt v2*trt
               v1*trt*center v2*trt*center / P PM;
  RANDOM intercept v1 v2 / SUBJECT=pid;
  MAKE 'PREDICTED' OUT=all;
  MAKE 'PREDMEANS' OUT=fixed;
RUN;
data random;
  merge
all(rename=(_pred_=all))fixed(rename=_pred_=fixed);
  random=all-fixed;
RUN;
```

includes both fixed and random effects. The random effects between the curly brackets represent a quadratic response curve for each individual patient. The analysis may be viewed as this: Each patient's responses are characterized with a quadratic curve, and the mean response curves are calculated from the individual response curves in the treatment groups in each center. The SAS code after the model produces both the average and individual response profiles.

PROC MIXED does not directly output estimates of random effects. They have to be derived from two data sets, PREDICTED and PREDMEANS. The data step after PROC MIXED computes the random effects.

A.5 PROC GENMOD for ANOVA on an arbitrary scale

PROC GENMOD allows the users to choose from a variety of mathematical distributions and scales for the analysis of variance. The following model uses the logarithmic scale:

$$\log[\text{mean}(\text{fev1})] = \text{baseline} + \text{center} + \text{treatment} + \text{center-treatment interaction},$$

and the following SAS code computes the summary measures:

Mean and Standard Error at Each Visit

```
PROC GENMOD;
  BY VISIT;
  CLASS center trt;
  MODEL fev1 = bsl center trt center*trt
               / pscale dist=poisson link=log
obstats;
  MAKE 'OBSTATS' out=obstats;
RUN;
```

PSCALE allows a scalar to be incorporated in the model to represent residual variations, and is indispensable in case that some mathematical distributions, like the Poisson and binomial, do not have independent parameters to represent the residual mean sum of squares. The DIST option specifies a mathematical distribution to represent the observed data

frequencies. For the analysis of variance, the normal distribution is most convenient regardless of the scale of choice. The LINK option specifies the scale. OBSTATS is a data set with the means and their standard errors.

If, besides a complex scale, a matrix is required to represent the residual mean sum of squares, not uncommon for the analysis longitudinal data, the REPEATED statement can be added. The analysis of variance technique with both complex scale and residual matrix is known as the generalized estimating equations (GEEs). The analysis specified in the model:

log[mean(fev1)] = baseline + center + visit + treatment + visit-treatment interaction + visit-treatment-center interaction

compares the effects of treatment across centers at each visit. If GEEs are mandatory, the following SAS code may be used to carry out the computation:

Mean and Standard Error at Each Visit

```
PROC GENMOD;
  CLASS pid visit center trt;
  MODEL fev1 = bsl center trt visit visit*trt
visit*trt*center
               / dist=normal link=log obstats;
  REPEATED SUBJECT=pid / TYPE=AR(1);
  MAKE 'OBSTATS' out=obstats;
RUN;
```

Unlike PROC MIXED where the EMPIRICAL option is required to direct the program computing the standard errors with the residuals, PROC GENMOD, at the time of writing this appendix, automatically utilizes residuals to estimate the standard errors. Therefore, the structure of the residual matrix is inconsequential. In this example, the first-order autoregressive structure is picked, specified as an option of the REPEAT statement, type = AR(1).

B Linear models for the analysis of variance

Linear models are the core computational technique for virtually all the analysis of variance techniques described in this book. In more mathematically oriented statistical textbooks, there are a variety of technical specifications with respect to parameterization and ways to derive normal equations for solving out parameters. I must emphasize that those specifications are strictly technical. The wording of two commonly seen technical specifications has caused a great deal of confusion, and that confusion is often exaggerated when inexperienced statisticians try to explain them to their non-statistical audience.

One technical specification pertains to the use of normal distribution to derive normal equations for resolving parameters. The wording of that specification often reads: "The data is assumed to be normally distributed," or "the data is assumed to follow a normal distribution." This specification by no means implies that the data must follow whatever a specified mathematical distribution in order for the analysis to be valid. It should be understood as this: The normal distribution is particularly convenient to compute the mean and its standard error because it is admissible to all data values and has two parameters with one representing the mean and the other the standard error.

The other technical specification pertains to the computation of pooled residual mean sum of squares. The wording of this specification often reads: "The variations of the data are assumed to be homogeneous." This specification must not be interpreted that the variations of the data in different groups have to be the same for the analysis to be valid. It is merely a technical stipulation to direct the mathematical manipulations so that the designated parameter, σ^2 for instance, represents the pooled residual mean sum of squares. A technical difficulty with complex scale is that the parameter designated to represent residual mean sum of squares functionally related to the parameter designed to represent the mean. With the Poisson distribution on logarithmic scale, for instance, if we use λ to represent the mean, the parameter representing its standard error is $\lambda\sigma^2$, where σ^2 is the overdispersion parameter. Given the sufficient statistics, λ and $\lambda\sigma^2$, although functionally related, are in fact independent. Once the mean has determined, its standard error is also determined and the residuals are the sufficient statistics, and it is really a matter of naming the

standard error with σ^2, $\lambda\sigma^2$ or Catherine. Since "homogeneity of variation" is just a rule of mathematical manipulation for our purpose, a check for this "homogeneity assumption" is catch 22, and entirely unnecessary.

Mathematics has no divine mysteries. It is nothing but human operations for human purposes.

B.1 The maximum likelihood technique

Let x_1, x_2, ..., x_n denote n observations, with their frequencies represented by $f(x_1, \theta)$, $f(x_2, \theta)$, ..., $f(x_n, \theta)$, where f is a convenient mathematical distribution with unknown parameters denoted by vector θ. The maximum likelihood estimate of θ is a set of values Θ such that $f(x_1, \Theta)$, $f(x_2, \Theta)$, ..., $f(x_n, \Theta)$ best represent the observed data frequencies *on average*. The process of maximization is done by forming the log-likelihood function,

$$L = \log(\lambda) = \log\left(\prod_{i=1}^{n} f(x_i, \theta)\right) = \sum_{i=1}^{n} \log f(x_i, \theta)$$

and then mathematically maximizing the log-likelihood function. The score equations, $\partial\mathbf{L}/\partial\theta$, are sufficient for estimating θ. $-\partial^2\mathbf{L}/\partial\theta^2$, known as the Fisher information, measures the precision of Θ, the maximum likelihood estimates (MLE) of θ. The inverse of Fisher information approximates the standard errors of the MLEs.

The frequency distribution of the data values is *not* required to resemble any mathematical distribution. The choice of mathematical distribution is technical. The minimal criteria for a mathematical distribution for this purpose are

- an adequate number of independent parameters to represent the desired summary measures, and
- the range of the mathematical function must cover all admissible data values.

For the analysis of variance, the normal distribution is convenient. It has two independent parameters, one representing the mean and the other the standard error. On the other hand, the Poisson and binomial distributions, for instance, are not adequate for that purpose, not only

because they are not admissible to negative values, but also because they do not have two independent parameters to represent the mean and its standard error. Nevertheless, if the data values are all positive, and we add a parameter, known as overdispersion parameter as what the PSCALE option does in PROC GENMOD, to represent the residual mean sum of squares, these two mathematical distributions can be used to carry out all the computations that we would normally do with a normal distribution.

The maximum likelihood estimates are *consistent* with the mathematical function that is used to represent the observed data frequencies. They are also *efficient* in the sense that the Fisher information approaches to the maximum with increasing number of patients. The Fisher information may be used independently to measure the precision of maximum likelihood estimates. The validity of this measure has nothing to do with the sample size. If p-value is required, one may multiple the MLEs with their Fisher information, and then *compare* this product to the standard normal distribution.

B.2 General linear models

Let $\mathbf{Y}$ denote the vector of responses and $\mathbf{X}$ denote the values of the explanatory variables. A general linear model is defined as $E(\mathbf{Y}) = \mathbf{X}\beta$, where E denotes expectation or averaging. The frequencies of $\mathbf{Y}$ are represented with normal distributions, $N(\mathbf{X}\beta, \mathbf{I}\sigma^2)$, where N denotes normal distribution, $\mathbf{X}\beta$ denotes the means, and $\mathbf{I}\sigma^2$ denotes the residual mean sum of squares. The maximum likelihood estimates of $E(\mathbf{Y})$ are

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y},$$

where "$-$" denotes generalized inverse, and the MLE of σ^2 is

$$[\mathbf{Y} - E(\mathbf{Y})]' [\mathbf{Y} - E(\mathbf{Y})]/n,$$

where n is the number of observations in $\mathbf{Y}$. The standard errors of the means are the diagonal elements in matrix

$$\text{var}(\mathbf{X}\beta) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\sigma^2.$$

B.3 Generalized linear models on an arbitrary scale

Let $\mathbf{Y}_i = (y_{i1}, y_{i2}, ..., y_{it})$ denote the t responses of patient i. A generalized linear model is defined as

$$g\,[\mu_i = E(\mathbf{Y}_i)] = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i, \quad i \text{ being any of the n patients,}$$

where g is the function representing the scale of choice. $\mathbf{X}_i\beta$ represents the mean responses in the group that patient i belongs to, and $\mathbf{Z}_i\mathbf{b}_i$ represents the average responses over the observations from patient i. $\mathbf{X}_i\beta$ is generally referred to as the fixed effects and $\mathbf{Z}_i\mathbf{b}_i$ the random effects. Random effects contribute to the variation of the mean responses represented by the fixed effects, and the contribution is denoted by

$$\text{var}(\mathbf{Z}_i\mathbf{b}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'.$$

The maximum likelihood estimation of β and $\mathbf{b}_i$'s goes through an iterative process. Function g is first linearized by the first-order Taylor series expansion:

$$\mathbf{U}_i = g(\mu_i) + \partial g(\mu_i)/\partial\,\mu_i\,(\mathbf{Y}_i - \mu_i) = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + g'(\mu_i)\,\mathbf{e}_i\,,$$
$$\text{for any of the n patients,}$$

where $\mathbf{e}_i = \mathbf{Y}_i - \mu_i$. The first two moments of $\mathbf{U}_i$ are

$$E(\mathbf{U}_i) = \mathbf{X}_i\beta \text{ and}$$

$$\text{var}(\mathbf{U}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \text{diag}[g'(\mu_i)]\,\text{var}(\mathbf{Y}_i)\,\text{diag}[g'(\mu_i)]\},$$

where $\text{diag}[g'(\mu_i)]$ is a diagonal matrix. Then a multivariate normal distribution is used to represent the frequency of $\mathbf{U}_i$:

$$N\,[E(\mathbf{U}_i), \text{var}(\mathbf{U}_i)]\,.$$

It is straightforward to form the log-likelihood function and apply the available numerical techniques to solve out β and $\mathbf{b}_i$'s. These β and $\mathbf{b}_i$'s are then plugged back into the Taylor series expansion formula to update

$\mathbf{U}_i$, followed by another iteration of the maximum likelihood estimation of β and $\mathbf{b}_i$'s. The entire process consists of linearization, maximization and updating (LIMU). This LIMU iteration goes on and on till the β and $\mathbf{b}_i$'s series converge.

If g is identity, the model degenerates to a mixed linear model:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \ , \quad \text{for any of the n patients.}$$

If the random terms are dropped out and the Fisher scoring algorithm is used to solve the likelihood function, the model reduces to the generalized estimating equations (GEEs):

$$\mathbf{U}_i = g(\mu_i) + \partial g(\mu_i)/\partial \mu_i \, (\mathbf{Y}_i - \mu_i) = \mathbf{X}_i\beta + g'(\mu_i)\, \mathbf{e}_i \ ,$$
$$\text{for any of the n patients.}$$

Of course, if g is identity, the random terms are dropped out, and var($\mathbf{Y}_i$) is represented with a scalar, the model is simply a general linear model:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{e}_i \ .$$

More technical details on LIMU were given in the dissertation of Xie.

C Analysis of variance for integrating a series of studies

The technique presented in this appendix is useful to perform an integrated analysis of variance on data from multiple studies that share the common treatment of interest. It is particularly useful when the studies to be integrated are designed differently with respect to group setup, stratification, and covariates. These differences dictate that the data from those studies cannot be simply pooled and analyzed with a single linear model.

The idea is this: First, global factors are separated from local factors. Global factors are those of shared interest in the studies to be integrated. Local factors are those specific to individual studies. Then, we use a linear model, with both local and global factors, to specify the analysis for each study. The efficient scores from the maximum likelihood estimation of the effects of global factor are used to represent the

contributions from individual studies. These global efficient scores are then summed up over studies, and the effects of the global factors are estimated from the combined efficient scores.

Let $\mathbf{Y}_i$ denote the vector of responses for study i, and $\mathbf{X}_i$ denote the values of the explanatory variables for the same study. The explanatory variables are partitioned into two sets: $\mathbf{X}_{i1}$ represents the variables that are specific for study i, and $\mathbf{X}_{i2}$ represents the variables that are shared by all the studies to be integrated. If we use the linear model,

$$E(\mathbf{Y}_i) = \mathbf{X}_{i1}\beta_{i1} + \mathbf{X}_{i2}\beta_2 ,$$

to specify the analysis for study i, the score equations for both β_{i1} and β_2 are

$$s(\beta_{i1}) = (\mathbf{X}_{i1}'\mathbf{X}_{i1})\,\beta_{i1} + \mathbf{X}_{i1}'(Y - \mathbf{X}_{i2}\beta_2) \text{ and}$$

$$s(\beta_2)_i = (\mathbf{X}_{i2}'\mathbf{X}_{i2})\,\beta_2 + \mathbf{X}_{i2}'(Y - \mathbf{X}_{i1}\beta_{i1}).$$

The combined estimate of β_2 is obtained by first adding up the score equations for β_2 from all the studies and then solving the combined score equations.

This idea can be simply extended to the analysis of variance on an arbitrary scale with the LIMU algorithm discussed in Appendix B. A generalized linear model may be specified for each of the p studies,

$$g\,[\mu_i = E(\mathbf{Y}_i)] = \mathbf{X}_{i1}\beta_{i1} + \mathbf{X}_{i2}\beta_2 , \quad i = 1, 2, ..., p,$$

where g represents the scale of choice. By linearization,

$$\mathbf{U}_i = g(\mu_i) + \partial g(\mu_i)/\partial\,\mu_i\,(\mathbf{Y}_i - \mu_i) = \mathbf{X}_{i1}\beta_{i1} + \mathbf{X}_{i2}\beta_2 + g'(\mu_i)\,\mathbf{e}_i ,$$

where $\mathbf{e}_i = \mathbf{Y}_i - \mu_i$, and the score equations for β_{i1} and β_2 are, respectively,

$$s(\beta_{i1}) = (\mathbf{X}_{i1}'\mathbf{W}\mathbf{X}_{i1})\,\beta_{i1} + \mathbf{X}_{i1}'\mathbf{W}(\mathbf{U}_i - \mathbf{X}_{i2}\beta_2) \text{ and}$$

$$s(\beta_2)_i = (\mathbf{X}_{i2}'\mathbf{W}\mathbf{X}_{i2})\,\beta_2 + \mathbf{X}_{i2}'\mathbf{W}(\mathbf{U}_i - \mathbf{X}_{i1}\beta_{i1}) ,$$

where

$$W^{-1} = \text{var}(\mathbf{U}_i) = \text{diag}[g'(\mu_i)]\ \text{var}(\mathbf{Y}_i)\ \text{diag}[g'(\mu_i)]\}.$$

The score equations for β_2 can be simply added up together, and the Fisher information of the combined β_2 is

$$I(\beta_2) = \sum_{i=1}^{p} \mathbf{X}_{i2}'\mathbf{W}\mathbf{X}_{i2}\,.$$

A simple algorithm is this: First, find out the maximum likelihood estimates for both β_{i1}'s and β_2's in all the studies. Then, define $\mathbf{Z}_i = \mathbf{U}_i - \mathbf{X}_{i1}\beta_{i1}$, and solve the combined score equation for β_2 iteratively. Given β_{i1}'s and β_2 , the residuals,

$$\mathbf{Y}_i - g^{-1}(\mathbf{X}_{i1}\beta_{i1} + \mathbf{X}_{i1}\beta_2),$$

can be used directly to estimate $\text{var}(\mathbf{Y}_i)$.

D The results of 18 trials on β blockade

Study	Drug	NYHA	Duration	N	Treatment	Death	Hospital	LVEF	STD
Engelmeier	Metoprolol	II-IV	12	16	placebo	2	4	22	14
Pollock	Bucindolol	III-IV	3	7	placebo	0	0	29	10
Lechat	Nebivolol	III-IV	1.5	6	placebo	0	0	.	.
Woodley	Bucindolol	II-IV	3	20	placebo	0	2	21	9
Mdc	Metoprolol	I-IV	13.2	189	placebo	21	49	28	12
Wisenbaugh	Nebivolol	II-III	3	13	placebo	0	0	23	9
Fisher	Metoprolol	III-IV	6	25	placebo	2	8	.	.
Bristow1	Bucindolol	II-III	3	34	placebo	2	3	29	8
Cibis	Bisoprolol	III-IV	12	321	placebo	67	82	25	11
Eichhorn	Metoprolol	II-III	3	9	placebo	0	2	17	8
Metra	Carvedilol	II-III	4	20	placebo	0	2	19	5
Olsen	Carvedilol	II-IV	4	23	placebo	0	0	20	10
Anz	Carvedilol	I-III	19.2	208	placebo	29	33	29	10
Krum	Carvedilol	II-IV	3.5	16	placebo	2	2	16	7
Bristow2	Carvedilol	II-IV	6.8	84	placebo	13	8	25	8
Packer	Carvedilol	II-IV	6.8	145	placebo	11	18	24	9
Colucci	Carvedilol	II-III	15	134	placebo	5	9	25	10
Cohn	Carvedilol	II-IV	8	35	placebo	2	1	23	9

The following algorithm is used to derive a single value for duration when the original data are a range: duration = 0.8 x lower end + 0.2 x upper end.

References

Fisher, RA. Statistical Methods for Research Workers, Seventh Edition. Edinburgh: Oliver & Boyd, Ltd. 1938.

Fisher, RA. The Design of Experiments, Seventh Edition. Edinburgh: Oliver & Boyd, Ltd. 1960.

Fisher, RA. Statistical Methods and Scientific Inference, Second Edition. Edinburgh: Oliver & Boyd, Ltd. 1959.

SAS Institute Inc., SAS/STAT® Software: Changes and Enhancements through Release 6.12, Cary, NC: SAS Institute, Inc. 1997.

SAS Institute Inc., SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2, Cary, NC: SAS Institute, Inc. 1989.

Belmont Research, Inc., CrossGraphs™ User Guide and Graph Types, Version 2. Cambridge, MA: Belmont Research, Inc. 1998.

Randel RK., Varley SA., Webb CF. Clinical Data Management. England: John Wiley & Sons, Ltd. 1993.

Boissel JP., Blanchard J., Panak E., Peyrieux JC., Sacks H. Considerations for the meta-analysis of randomized clinical trials: summary of a panel discussion. Controlled Clinical Trials 10:254-281. 1989.

Lechat P., Packer M., Chalon S., Cucherat M., Arab T., Boissel JP. Clinical effects of β-adrenergic blockade in chronic heart failure - a meta-analysis of double-blind, placebo-controlled, randomized trials. Circulation. 98:1184-1191. 1998.

Cox DR. and Oakes D. Analysis of Survival Data. London: Chapman and Hall. 1984.

Xie X. A generalized mixed linear model for the analysis of longitudinal data. A dissertation in the Louisiana State University Medical Center Library, New Orleans. 1995.

Neyman J. and Pearson ES. Joint Statistical Papers. Cambridge University Press. 1967.

Hedges LV, and Olkin I. Statistical Methods for Meta-analysis. New York: Academic Press. 1985.

DerSimonian R. and Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 7:177-188. 1986.

Index

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W