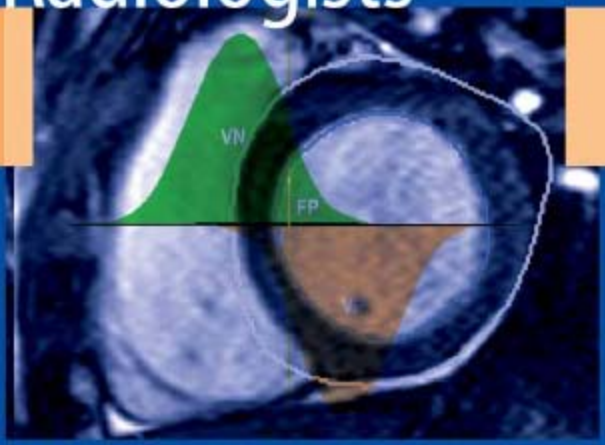
Francesco Sardanelli
Giovanni Di Leo

Biostatistics for Radiologists



Planning, Performing, and Writing a Radiologic Study

Foreword by Adrian K. Dixon

Springer

Biostatistics for Radiologists

Planning, Performing, and Writing a Radiologic Study

Francesco Sardanelli • Giovanni Di Leo

Biostatistics for Radiologists

Planning, Performing, and Writing a Radiologic Study

Foreword by
Adrian K. Dixon

Springer

FRANCESCO SARDANELLI, MD
Associate Professor of Radiology
University of Milan School of Medicine
Department of Medical and Surgical Sciences
Director of Radiology Unit
IRCCS Policlinico San Donato
San Donato Milanese, Italy
e-mail: francesco.sardanelli@unimi.it

GIOVANNI DI LEO, Dr.Sci.
Researcher
Radiology Unit
IRCCS Policlinico San Donato
San Donato Milanese, Italy
e-mail: gianni.dileo77@gmail.com

Originally published as:
Biostatistica in Radiologia
Progettare, realizzare e scrivere un lavoro scientifico radiologico
Francesco Sardanelli, Giovanni Di Leo

Library of Congress Control Number: 2008938920

ISBN 978-88-470-1132-8 Springer Milan Berlin Heidelberg New York
e-ISBN 978-88-470-1133-5

Springer is a part of Springer Science+Business Media
springer.com

Typesetting: Ferrari – Studio editoriale, Cologno Monzese (MI), Italy
Printing and binding: Printer Trento S.r.l., Trento, Italy

Printed in Italy
Springer-Verlag Italia S.r.l., Via Decembrio 28, I-20137 Milan, Italy

To Genny, Francesca, and Federica
F.S.

To my father,
who taught me how to distinguish
significant from nonsignificant things
G.D.L.

Foreword

It is a pleasure to write the foreword to this excellent book on the evidence behind radiological investigations and biostatistics in radiology. This is an area which is not widely appreciated by radiologists and it would be an invaluable book for those in training who should become fully versed about terminology such as technical performance, diagnostic performance, diagnostic impact, therapeutic impact, patient impact, patient outcomes and societal impact. They should also know that the widely (and often erroneously) used term 'accuracy' may not be the best assessment!

This very readable book should not only be useful for radiologists but also administrators who are now beginning to realise that an effective imaging department underpins all high quality cost-efficient modern medicine. The history and development of evidence-based radiology (from Fryback and Thornbury's original paper right up to date with recent contributions from Hollingworth, Hunink, Jarvik and Malone) is very well presented.

Lorenzo Mannelli, an excellent Italian Radiologist working in Cambridge, says of the Italian Edition: "The book is easy to read and the short paragraph titles on the side of the pages make it easy to use for future reviewing of "hot topics" when needed. All the examples and vocabulary are from the radiological world, making the statistics easier to understand. Although after reading this book, you will still need statistical advice, at least you will be able to understand what the statistician is speaking about! The final chapter on impact factors is interesting and helps the reader to understand the dynamics of journals. I definitely recommend this book as an easy reading for residents in radiology".

I am sure that this new English language edition will fill a very major void in the radiological literature. The authors have done us a great service.

Cambridge, UK, October 2008 *Adrian K. Dixon, MD*

Preface to the Italian Edition

Better to light a candle
than to curse the darkness.

CONFUCIUS

Science deals with discovery
but also with communication.
It's hard to say you have an idea
if you are not able to evoke the same idea
in the mind of your listener.

MARCUS DU SATOY

For many years "Biostatistics for Radiologists" was an unrealized dream of the senior author. Since that dream is now coming true, I have taken on the task of writing this preface, which offers the opportunity for an appraisal of the years leading up to the genesis of this book. I hope it can be useful to young colleagues who intend to devote themselves to radiologic research.

More than twenty-five years ago, I was a resident at the Postgraduate School in Radiodiagnostics of the University of Genoa, directed by Professor Luigi Oliva. My supervisor was Professor Giorgio Cittadini, Director of the Chair "R" of Radiology of the University. He was the chairman of my medical graduation thesis, entitled "Colonic hypotonic effect of fenoverine and hyoscine N-butyl bromide: analysis of variance with nonparametric tests". Already then there was an attention to statistical methods predicting the events of my future before me.

In 1984, after several years mainly dedicated to gastrointestinal double-contrast studies, I was included in a small team made up of physicians, physicists, and engineers who had the good fortune to work on one of the first magnetic resonance imaging scanners installed in Italy. It was a prototype with a resistive magnet operating at only 0.15 T. For the best use of this new diagnostic technology, the radiologist had to understand the NMR phenomenon, the radiofrequency pulse sequences, and the role of the field gradients which generate the images. At that time, physicists were teaching magnetic resonance in courses and congresses using formal demonstrations based on Bloch's equations, combining both classic and quantum models. These lessons were very hard to follow. Only when formulas and equations were translated into a different language, positively associated with the clinical meaning of the images, did the attending radiologists open their eyes and grasp the practical sense of that theory. For the first time I understood that scientific communication was a crucial process requiring intellect, fantasy, and creativity.

In the same year, I was involved as author in a small paper entitled "Sensitivity, specificity, overall accuracy. What is the meaning of these three words commonly used in scientific radiologic language?"[1]. This paper was the result of an interesting discussion which had begun in Cittadini's room late one evening and lasted for two-three hours. The topic – how to quantify diagnostic performance – seemed highly intriguing to me. I promised myself to gain a deeper insight into the matter. It was a new world waiting to be explored: how to evaluate the uncertainty intrinsic to biologic phenomena and measurements and, as a consequence, medical diagnosis. At the time I was only a resident cooperating towards writing an article, but I began to add some substance to the immaterial dream of writing a book. The chapter dedicated to "Indices of Diagnostic Performance" included in the Italian textbook "Diagnostic Imaging and Radiotherapy", recently published in its sixth edition, would be the embryonic stage of "Biostatistics for Radiologists".

Some years later, in 1987, I became a staff radiologist of the Chair R of Radiology at the Genoa University and San Martino Hospital. I began to add a planned research activity to the clinical routine. The areas of research involving high-level cooperation with clinicians produced the most interesting results. This was the case with Giuseppe Molinari (cardiologist at the Genoa University) and Giuseppe Canavese, breast surgeon at the Genoa Cancer Research Institute. In this period I submitted my first manuscripts to peer-reviewed journals. Immediately, I understood that good technical knowledge and updated clinical experience are not enough for writing an article enough good to be accepted for publication. The crux of the matter is given by study design, data presentation and analysis, and, in particular, statistical methods which demonstrate the significance of the results.

Around this time I began to interact with statisticians. Once again communication was stifled. Radiologists were on one side of a wall and statisticians on the other, as it had been earlier with the physicists for magnetic resonance. In my personal experience, however, this wall crumbled thanks to research conduced on multiple sclerosis. In this field, magnetic resonance imaging was playing an increasingly important role. My relations with Gianluigi Mancardi (Department of Neurology, University of Genoa) and Paolo Bruzzi (Clinical Epidemiology, Genoa Cancer Research Institute) were a turning point. Each of us wanted to learn what the other two colleagues already knew and would spend hours and hours to reach… understanding.

At the same time, another apparently impersonal factor was in action – the reviewers analyzing the manuscripts I submitted to the journals. Their criticisms were sometimes very harsh. However, the higher the rank of the journal, the greater the knowledge in methodology I could obtain by interacting with the reviewers, even though the manuscript was rejected. This was another way I begun to accumulate a limited know-how in biostatistics and research methodology applied to radiology. While I was (and still am) learn-

[1] Sardanelli F, Garlaschi G, Cittadini G (1984) Sensibilità, specificità, accuratezza diagnostica. Quale significato attribuire a queste tre parole così spesso usate nel linguaggio scientifico radiologico? Il Radiologo 23:58-59.

ing from my errors, a long line of textbooks on medical statistics began to grow in my bookcase.

At the beginning of the 1990s, I became head of Diagnostic Imaging at the Breast Unit of the San Martino University Hospital and the Genoa Research Cancer Institute. By that time, Italian breast radiologists were involved in a flourishing debate: clinical mammography on the one hand, organized screening mammography on the other. Which was the key-point? The majority of Italian women who asked for a mammographic examination in radiology departments were asymptomatic. Their request was one of spontaneous periodic control. This gave rise to a kind of oxymoron: clinical mammography (with physical examination and frequent accessorial ultrasound examination) in an asymptomatic population. Radiologists with lengthy clinical experience of breast imaging on patients with symptoms were conditioned by a practice of "first of all, sensitivity" which attained acceptable levels of specificity with further work-up in a relevant fraction of patients. The application of this logic to asymptomatic women resulted in the medicalization of a healthy population. It was a typical problem generated by low disease prevalence. In asymptomatic women sensitivity must be combined with a sufficiently high specificity and positive predictive value. On the other hand, we knew that ultrasound enabled us to detect breast cancers in women with high breast density and negative mammography and that periodical mammography is also useful in women under 50 and over 70. However, once again there was a wall. Clinical mammography on one side, screening mammography on the other. This experience gave me new incentive to flesh out that dream.

In the meantime, I begun to serve as a reviewer for international journals. This gave me the opportunity to compare my evaluation of a manuscript with those of other reviewers. Moreover, at the end of that decade I started to cooperate with Franca Podo from the Istituto Superiore di Sanità (Rome), a physicist and world-renowned expert in magnetic resonance imaging and spectroscopy. Working together we conducted the HIBCRIT study for multimodality surveillance of women at high genetic-familial risk of breast cancer. Here the high disease prevalence justified an intensive surveillance including physical examination, mammography, ultrasound, and contrast-enhanced magnetic resonance imaging. It was a fantastic experience from which I learned a lot, especially on the management of multicenter trials, an intense and effective cooperation without the need of breaking down walls, with a follow-on now extending to new topics.

From 1999 to 2000 I was Director of the Department of Radiology at the Biomedical Institute in Genoa. This role broadened the spectrum of my experience. The higher levels of productivity in clinical radiologic activity was a preparation for the upcoming events.

In fact, in 2001 I was assigned the Direction of the Department of Radiology at the Policlinico San Donato near Milan. In my opinion, the principal aim was to have a radiologic team with high levels of clinical efficiency and scientific research. The administrators gave me free reign to start a process of training and selection of young colleagues. Some of the fruits of this process can already be appreciated. I have to thank several persons who have been and continue to be keystones in the day-to-day operation of the system: the radiologists Alberto Aliprandi, Bijan Babaei and Pietro Bertolotti and the coordinators of

radiographers Francesco Gerra and Eleonora Norma Lupo. Recently we were joined by Carlo Ottonello, who was resident in Radiodiagnostics at the Genoa University at the beginning of the 1990s. Our younger colleagues have the opportunity to show their abilities in clinics and research, in part thanks to many projects we have in cooperation with the clinical departments of our institution.

In recent years, I combined the Direction of the Unit of Radiology of the Policlinico San Donato (from 2006, appointed as Istituto di Ricovero e Cura a Carattere Scientifico, IRCCS, by the Ministry of Health) with the position of Associate Professor of Radiology at the University of Milan School of Medicine. This new context favored my study on research methodology. The last chapter of this book arose from a lesson entitled "How to Write a Scientific Paper?". I held with the residents of the Postgraduate School in Radiodiagnostics on the express request of the Director of the School, Professor Gianpaolo Cornalba, in a framework of close cooperation and common rationale.

At the same time I served on the National Board of the Councilors of the Italian Society of Medical Radiology as a President's Delegate for Scientific Research. Over the past four years, Alessandro Del Maschio (at that time President's Delegate for Scientific Research) promoted a course on Methodology of Scientific Research. It was held by Irene Floriani and Valter Torri, from the "Mario Negri" Institute (Milan) and then repeated in several Italian cities. The aim was to increase the level of knowledge in research methodology among Italian radiologists, a need which had already emerged during the first multicenter studies promoted by the Italian Society of Medical Radiology (SIRM) on breast MR imaging. My involvement only enlarged the scale of the audience by introducing several radiologists and a young physicist (the second author of this book) to the faculty. We all worked together in the preparation of the lessons during multiple meetings and long discussions, in particular with Irene Floriani and Valter Torri. This was a new stimulus for realizing my dream. So they too deserve my heartfelt thanks.

However, something was still missing: I had no solid mathematical background. Giving prominence to logics over computing could not exempt me from formal correctness. I therefore decided to associate a clever physicist from the Naples School and full-time researcher at the Radiology Unit of the IRCCS Policlinico San Donato, Giovanni Di Leo, with the project of this book. Both of us worked on all the chapters, even though he drafted the first version of the chapters with a higher mathematical content while I drafted the first version of the chapters with a higher logical and methodologic content, with each of us providing the other with constructive criticism.

Lastly, I would like to thank Antonella Cerri from Springer. She enthusiastically latched on to the idea of this book when I described the project to her several years ago during a friendly chat at the end of a meeting of the editorial board of European Radiology.

We really hope to communicate to radiologists the methodologic know-how which is taking on an increasingly important role. Many years ago a surgeon asked me: "Do you know what the difference is between a radiologist and a surgeon?". Before I could answer, he said: "You say 'my CT scanner', I say 'my patient'". It was true. We must give clear demonstrations that the images

of higher and higher quality we are able to produce have a significant impact on patient outcome and the population health status.

This book is a small contribution towards this challenge.

As I stated at the beginning of this preface, younger colleagues could heed the advice from this personal history. When you wake up in the morning, keep on dreaming. Then sooner or later, flesh out those dreams and bring them to life.

San Donato Milanese, April 2008 *Francesco Sardanelli*

Preface to the English Edition

We enthusiastically accepted the proposal from Springer to do an English version of this book, based on the advantage that radiologists (and more generally experts in medical imaging) were unable to find a volume where the basics of research methodology were presented as applied to diagnostic imaging. Conversely, we had the large disadvantage related to the difficulty of explaining complex matters such as biostatistics which have been extensively developed in many other splendid books written by real experts in the field.

However, insofar as we went ahead in rewriting the text in English, we realized that not only was the meaning retained, but the message also became clearer and less redundant. This was probably due not only to the effect of the different language, but also the result of a rethinking of the content of chapters and paragraphs several months after publishing the Italian edition. Now, our general impression is that the Italian version has been written for ourselves (to hone our thinking, to render it more analytic and detailed, to better understand the subject matter) and that the English version has been written for the reader (to provide her/him with a clearer message). We hope that this is true. Obviously, small errors and imperfections have been corrected and some points specifically written for Italian radiologists have been omitted.

The major change made in this English version is an expanded Introduction with more emphasis on evidence-based medicine and evidence based radiology.

At any rate, we would like to emphasize that this book is nothing more than an introduction to the topic, a portal to the realm of research methodology, with the words "radiology and medical imaging" emblazoned upon it.

A sincere word of thanks to Alexander Cormack, the English copyeditor who had the patience to transform our text into real English.

San Donato Milanese, October 2008

Francesco Sardanelli
Giovanni Di Leo

Acknowledgements

A sincere word of thanks to:
– all the faculty members of the SIRM Course on Scientific Methodology, who cooperated with Irene Floriani, Valter Torri, and the two authors: Giuseppe Brancatelli, Laura Crocetti, Antonella Filippone and Roberto Carlo Parodi;
– Lorna Easton who searched for the impact factors of the medical journals reported in the tables in Chapter 10;
– Dr. Francesco Secchi for the systematic evaluation of the instructions for authors of the radiologic journals;
– Dr. Myriam G. Hunink (Erasmus University Medical Center, Rotterdam, The Netherlands) for the suggestions used in the Introduction to this English version;
– the authors and publishers who gave permission for the reproduction of tables and figures from previously published papers;
– Elisabetta Ostini for the layout of text, tables, and figures.

Contents

Introduction

The practice of evidence-based medicine means
integrating individual clinical expertise
with the best available external evidence
from systematic research.

DAVE L. SACKETT

The creative principle of science
resides in mathematics.

ALBERT EINSTEIN

After all, to understand
is the intrinsic purpose of science,
and science is really much more
than mechanical computing.

ROGER PENROSE

Evidence-Based Medicine (EBM)

Over the past three decades, the following view has gained increasing favor throughout the medical community: clinical practice should be based on the critical evaluation of the results obtained from medical scientific research. Today this evaluation is greatly favored by Internet which provides instantaneous online access to the most recent studies even before they appear in print form. The possibility of instantaneously accessing quality-filtered and relevance-filtered secondary publications (meta-analyses, systematic reviews, and guidelines) has become real in routine practice.

Evidence-based medicine (EBM)

This notion – a clinical practice based on the results (the *evidence*) given by the research – has engendered a discipline: *evidence-based medicine* (EBM), also referred to as *evidence-based healthcare*, or *evidence-based practice* [MALONE, 2007]. In this context the term *evidence* is more closely associated with the concepts of *proof*, *demonstration,* or *testability* than simply with *visibility* or *clarity*. In fact, the general meaning of the new discipline suggests a clinical practice no longer based on bequeathed knowledge, on opinions, impressions, and perceptions, but on demonstrable proofs. EBM has been defined as "the systematic application of the best evidence to evaluate the available options and decision making in clinical management and policy settings", i.e. "integrating clinical expertise with the best available external clinical evidence from research" [EVIDENCE-BASED RADIOLOGY WORKING GROUP, 2001].

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

Origins of EBM

This concept is not new. The basis for this way of thinking was developed in the 19th century (Pierre C.A. Luis) and during the 20th century (Ronald A. Fisher, Austin Bradford Hill, Richard Doll, and Archie Cochrane). However, it was not until the second half of the last century that the Canadian School led by Gordon Guyatt and Dave L. Sackett at McMaster University (Hamilton, Ontario, Canada) promoted the tendency to guide clinical practice using the best results – the evidence – produced by scientific research [Evidence-Based Radiology Working Group, 2001; Greenhalgh, 2006a]. This approach was subsequently refined also by the Center for Evidence-Based Medicine (CEBM) at University of Oxford, England [Centre for Evidence-Based Medicine (http://cebm.net); Malone, 2007].

Dave L. Sackett and coworkers stated that:

EBM definitions

Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external evidence from systematic research [Sackett et al, 1996].

A highly attractive alternative but more technical definition, explicitly including diagnosis and investigation, has been proposed by Anna Donald and Trisha Greenhalgh:

Evidence-based medicine is the use of mathematical estimates of the risk, of benefit and harm, derived from high-quality research on population samples, to inform clinical decision making in the diagnosis, investigation or management of individual patients [Greenhalgh, 2006b].

Patient's values and choice

However, EBM is not only the combination of current best available external evidence and individual clinical expertise. A third factor must be included in EBM: the patient's values and choice. "It cannot result in slavish, cookbook approaches to individual patient care" [Sackett et al, 1996]. Thus, EBM is the integration of: (i) research evidence; (ii) clinical expertise; and (iii) patient's values and preferences [Sackett et al, 1996; Hunink et al, 2001; Malone and Staunton, 2007]. Clinical expertise "decides whether the external evidence applies to the individual patient", evaluating "how it matches the patient's clinical state, predicament, and preferences" [Sackett et al, 1996]. A synopsis of this process is given in Figure 0.1.

Two general methods are generally proposed for applying EBM [Dodd, 2007; Malone and Staunton, 2007; van Beek and Malone, 2007] (Figure 0.2):

Top-down EBM

– the *top-down* method, when academic centers, special groups of experts on behalf of medical bodies, or specialized organizations (e.g. the Cochrane collaboration; http://www.cochrane.org) provide high-quality primary studies (original research studies), systematic reviews and meta-analyses, applications of decision analysis, or issue evidence-based guidelines and make efforts to put them into practice;

Bottom-up EBM

– the *bottom-up* method, when practitioners or other physicians working in routine practice are able "to ask a question, search and appraise the literature, and then apply best current evidence in a local setting", opening a so-called *audit cycle*.

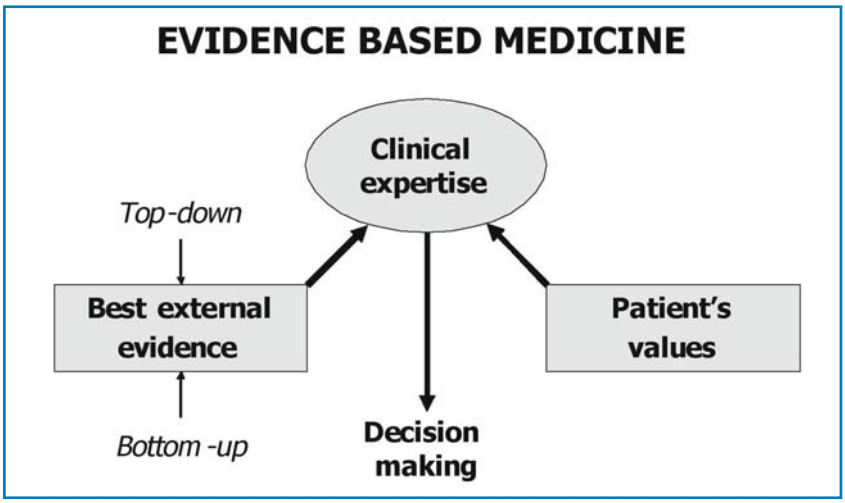


Figure 0.1. The general scheme of evidence based medicine. See Figure 0.2 for the top-down and bottom-up approaches to the best external evidence.

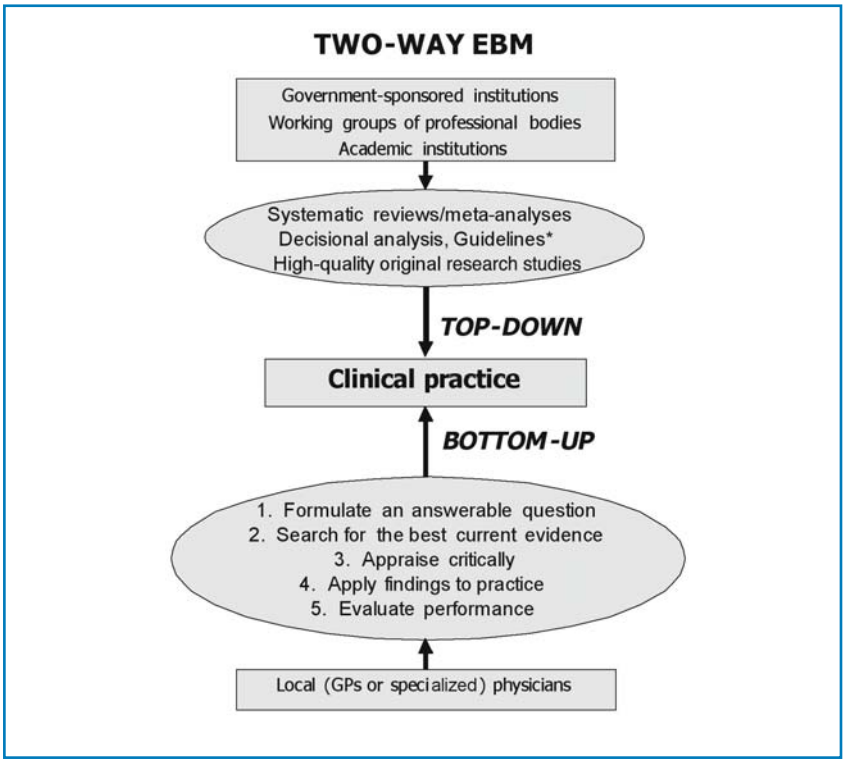


Figure 0.2. Top-down and bottom-up processes for evidence based medicine.
*Appropriateness criteria are not included in the top-down EBM method since they are based on expert opinion, even though formalized procedures (such as the Delphi protocol) are frequently used and experts commonly base their opinion on systematic reviews and meta-analyses [Medina and Blackmore, 2007].
EBM = Evidence Based Medicine.

We should note that the top-down method involves a small number of people considered experts and does not involve physicians acting at the local level. However, there is a difference between the production of systematic reviews and meta-analyses (which are welcome as an important source of information by local physicians who want to practice the bottom-up model) and the production of guidelines which could be considered as an external cookbook (mistaken for a mandatory standard of practice) by physicians who feel themselves removed from the decision-making process [VAN BEEK AND MALONE, 2007]. On the other hand, the bottom-up method (which was considered an EBM method before the top-down method [HOLLINGWORTH AND JARVIK, 2007]) implies a higher level of knowledge of medical research methodology and EBM techniques by local physicians than that demanded by the top-down method. In either case, a qualitative improvement in patient care is expected. At any rate, clinical expertise must play a pivotal role as integrator of external evidence and patient's values and choice. When decision analyses, meta-analyses and guidelines provide only part of the external evidence found by the local physicians, the two models act together, as hopefully should happen in practice. Moreover, a particular aim of the top-down method is the identification of gaps in knowledge to be filled by future research. In this way, EBM becomes a method for redirecting medical research towards improved medical practice [HOLLINGWORTH AND JARVIK, 2007].

EBM limitations

However, EBM is burdened by limitations and beset by criticisms. It has been judged as unproven, very time-consuming (and therefore expensive), narrowing the research agenda and patients' options, facilitating cost cutting, threatening professional autonomy and clinical freedom [SACKETT ET AL, 1996; TRINDER, 2000; MALONE AND STAUNTON, 2007]. At an objective evaluation, these criticisms seem to be substantially weak due to the pivotal role attributed to "individual clinical expertise" by EBM and to the general EBM aim "to maximize the quality and quantity of life for the individual patient" which "may raise rather than lower the cost of their care" as pointed out by Dave L. Sackett in 1996 [SACKETT ET AL, 1996].

Other limitations seem to be more relevant. On the one hand, large clinical areas – radiology being one of them – have not been sufficiently explored by studies according to EBM criteria. On the other hand, real patients can be totally different from those described in the literature, especially due to the presence of comorbidities, making the conclusions of clinical trials not directly applicable. This event is the day-to-day reality in geriatric medicine. The ageing population in Western countries has created a hard benchmark for EBM. These limitations may be related to a general criticism which suggests that the central feature in the EBM perspective is the patient population and not the individual patient [TONELLI, 1998; RAYMOND AND TROP, 2007]. Lastly, we should avoid unbridled enthusiasm for clinical guidelines, especially if they are issued with questionable methods [WOOLF ET AL, 1999].

However, all these limitations appear more as problems due to a still limited development and application of EBM than intrinsic EBM limitations. Basically, the correctness of EBM should be borne in mind, in that EBM aims to provide the best choice for the individual real patient with the use of probabilistic reasoning. EBM is investing significant effort towards improving contemporary medicine.

Delayed Diffusion of EBM in Radiology and Peculiar Features of Evidence-Based Radiology

Evidence-based radiology (EBR)

Radiology is not outside of EBM, as stated by David L. Sackett and coworkers in 1996: "EBM is not restricted to randomised trials and meta-analyses […]. To find out about the accuracy of a diagnostic test, we need to find proper cross sectional studies of patients clinically suspected of harboring the relevant disorder, not a randomised trial" [Sackett et al, 1996]. *Evidence-based radiology* (EBR), also called *evidence-based imaging*, first appeared in the literature only in recent years.

EBR delay

Until 2000, few papers on EBR were published in nonradiologic journals [Acheson and Mitchell, 1993; No authors listed (British Columbia Office of Health Technology Assessment), 1997; No authors listed, Int J Assess Health Care, 1997; Dixon, 1997; Mukerjee, 1999] and in one journal specialized in dentomaxillofacial radiology [Liedberg et al, 1996]. From 2001 to 2005, several papers introduced the EBM approach in radiology [Evidence-Based Radiology Working Group, 2001; Taïeb and Vennin, 2001; Arrivé and Tubiana, 2002; Bui et al, 2002; Guillerman et al, 2002; Kainberger et al, 2002; Bennett, 2003; Blackmore, 2003; Cohen et al, 2003; Goergen et al, 2003; Medina et al, 2003; Blackmore, 2004; Dodd et al, 2004; Erden, 2004; Gilbert et al, 2004; Matowe and Gilbert, 2004; Giovagnoni et al, 2005]. Not until 2006 was the first edition of the book entitled *Evidence-Based Imaging* published by L. Santiago Medina and C. Craig Blackmore [Medina and Blackmore, 2006]. The diffusion of EBM in radiology was delayed. From this viewpoint, radiology is "behind other specialties" [Medina and Blackmore, 2007].

As a matter of fact, according to L. Santiago Medina and C. Craig Blackmore, "only around 30% of what constitutes 'imaging knowledge' is substantiated by reliable scientific inquiry" [Medina and Blackmore, 2006]. Other authors estimate that less than 10% of standard imaging procedures is supported by sufficient randomized controlled trials, meta-analyses or systematic reviews [Dixon, 1997; RCR Working Party, 1998; Kainberger et al, 2002].

Particular traits of radiology

The EBR delay may also be linked to several particular traits of our discipline. In fact, the comparison between two diagnostic imaging modalities is markedly different from the well-known comparison between two treatments, typically between a new drug and a placebo or standard care. Thus, the classic design of the randomized controlled trial is not the standard for radiologic studies. What are the peculiar features of radiology which need to be considered?

Tecnical expertise

First of all, the evaluation of the diagnostic performance of imaging modalities must be based on a deep insight of the technologies used for image generation and postprocessing. Technical expertise has to be combined with clinical expertise in judging when and how the best available external evidence can be applied in clinical practice. This aspect is just as important as "clinical expertise" (knowledge of indications for an imaging procedure, imaging interpretation and reporting, etc). Dodd and coworkers showed the consequences of ignoring a technical detail such as slice thickness in evaluating the diagnostic performance of magnetic resonance (MR) cholangiopancreatography: using a 3-mm instead of a 5-mm thickness, the diagnostic performance for the detection of choledocholithiasis changed from 0.57 sensitivity and 1.0 specificity to

0.92 sensitivity and 0.97 specificity [Dodd et al, 2004]. If the results of technically inadequate imaging protocols are included in a meta-analysis, the consequence will be the underestimation of diagnostic performance.

At times progress in clinical imaging is essentially driven by the development of new technology, as was the case for imaging at the beginning of the 1980s. However, more frequently, an important gain in spatial or temporal resolution, in signal-to-noise or contrast-to-noise ratio is attained through hardware and/or software innovations in pre-existing technology. This new step broadens the clinical applicability of the technology, as was the case for computed tomography (CT) which evolved from helical single-slice to multidetector row scanners, thus opening the way to cardiac CT and CT angiography of the coronary arteries. Keeping up to date with technologic development is a hard task for radiologists, and a relevant part of the time not spent with imaging interpretation should be dedicated to the study of new imaging modalities or techniques. For radiologic research, each new technology appearing on the market should be tested with studies on its technical performance (image resolution, etc.).

Reproducibility

Second, we need to perform studies on the reproducibility of the results of imaging modalities (intraobserver, interobserver, and interstudy variability), an emerging research area which requires dedicated study design and statistical methods (e.g. Cohen k statistics and Bland-Altman analysis). In fact, if a test shows poor reproducibility, it will never provide good diagnostic performance, i.e. sensitivity and specificity. Good reproducibility is a necessary (but not sufficient) condition for a test to be useful.

High speed of technologic evolution

Third, the increasing availability of multiple options in diagnostic imaging should be taken into consideration along with their continuous and sometimes unexpected technologic development and sophistication. Thus, the high speed of technologic evolution has created not only the need to study theory and practical applications of new tools, but also to repeatedly start with studies on technical performance, reproducibility, and diagnostic performance. The faster the advances in technical development, the more difficult it is to do the job in time. This development is often much more rapid than the time required for performing clinical studies for the basic evaluation of diagnostic performance. From this viewpoint, *we are always too late with our assessment studies*.

From images to patients

However, the most important problem to be considered with new diagnostic technology is that "a balance must be struck between apparent (e.g. diagnostic) benefit and real benefit to the patient" [Dixon, 1997]. In fact, a qualitative leap in radiologic research is now expected: from the demonstration of the increasing ability to see more and better, to the demonstration of a significant change in treatment planning or, at best, a significant gain in patient health and/or quality of life – the patient outcome.

The ALARA principle

Lastly, we should specifically integrate a new aspect in EBR, i.e. the need to avoid unnecessary exposure to ionizing radiation, according to the *as low as reasonably achievable* (ALARA) principle [No authors listed, Proceedings of the Second ALARA Conference, 2004; Prasad et al, 2004; Semelka et al, 2007] and to governmental regulations [Council of the European Union, 1997; Barr et al, 2006; FDA Radiological Health Program, 2008]. The ALARA principle might be considered as embedded in radiologic "technical and clinical expertise". However, in our opinion, it should be regarded as *a*

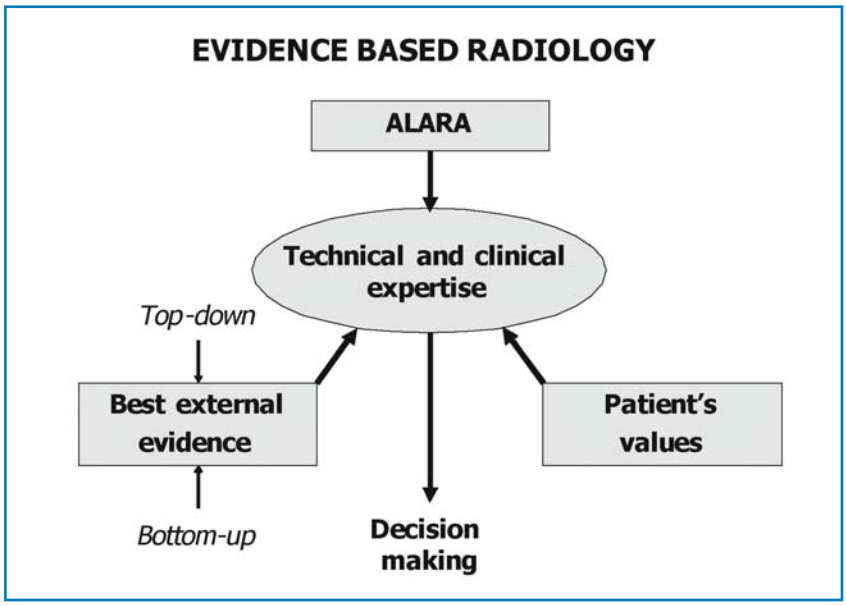


Figure 0.3. The process of evidence based radiology. ALARA = "as low as reasonably achievable", with reference to ionizing radiation exposure.

fourth dimension of EBR, due to the increasing relevance of radioprotection issues in radiologic thinking and practice. A graphical representation of the EBR process, including the ALARA principle, is provided in Figure 0.3.

Health Technology Assessment in Radiology and Hierarchy of Studies on Diagnostic Tests

Health technology assessment (HTA) in radiology

In the framework described above, EBM and EBR are based on the possibility of obtaining the best external evidence for a specific clinical question. Now the question is: how is this evidence produced? In other words, which methods should be used to demonstrate the value of a diagnostic imaging technology? This field is what we name *health technology assessment* (HTA) and particular features of HTA are important in radiology. Thus, EBR may exist only if a good radiologic HTA is available. As stated by William Hollingworth and Jeffry G. Jarvik, "the tricky part, as with boring a tunnel through a mountain, is making sure that the two ends meet in the middle" [Hollingworth and Jarvik, 2007].

According to the United Kingdom HTA Programme, HTA should answer the following four fundamental questions on a given technology [White et al, 2000; Hollingworth and Jarvik, 2007]:

1. does it work?
2. for whom?
3. at what cost?
4. how does it compare with alternatives?

Efficacy, effectiveness and efficiency

In this context, increasing importance has been gained by the use of three different terms. While *efficacy* reflects the performance of medical technology under ideal conditions, *effectiveness* evaluates the same performance under ordinary conditions, and *efficiency* measures the cost-effectiveness [Hillman and Gatsonis, 2008]. In this way the development of a procedure in specialized or academic centers is distinguished by its application to routine clinical practice and from the inevitable role played by the economic costs associated with implementation of a procedure.

Hierarchy of studies on diagnostic tests

To evaluate the impact of the results of studies, i.e. the level at which the HTA was performed, we need a hierarchy of values. Such a hierarchy has been proposed for diagnostic tests and also accepted for diagnostic imaging modalities. During the 1970s, the first classification proposed five levels for the analysis of the diagnostic and therapeutic impact of cranial CT [Fineberg et al, 1977]. By the 1990s [Fryback and Thornbury, 1991], this classification had evolved into a six-level scale, thanks to the addition of a top level called *societal impact* [Thornbury, 1994; Mackenzie and Dixon, 1995; Thornbury, 1999]. A description of this scale was presented more recently in the radiologic literature [Evidence-Based Radiology Working Group, 2001; Sunshine and Applegate, 2004].

The six-level scale

This six-level hierarchy scale (Table 0.1) is currently widely accepted as a foundation for HTA of diagnostic tools. This framework provides an opportunity to assess a technology from differing viewpoints. Studies on technical performance (level 1) are of key importance to the imaging community and the evaluation of diagnostic performance and reproducibility (level 2) are the basis for adopting a new technique by radiologists and clinicians. However, radiologists and clinicians are also interested in how an imaging technique impacts patient management (levels 3 and 4) and patient outcomes (level 5) while healthcare providers wish to ascertain the costs and benefits of reimbursing a new technique, from a societal perspective (level 6). Governments are mainly concerned about the societal impact of new technologies in comparison to that of other initiatives they may be considering.

A one-way logical chain

Note that this hierarchical order is a one-way logical chain. A positive effect at any level generally implies a positive effect at all preceding levels but not vice versa [Hollingworth and Jarvik, 2007]. While a new diagnostic technology with a positive impact on patient outcome probably has a better technical performance, higher diagnostic accuracy, etc. compared with the standard technology, *there is no certainty that a radiologic test with a higher diagnostic accuracy results in a better patient outcome*. If we have demonstrated an effective diagnostic performance of a new test (level 2), the impact on a higher level depends on the clinical setting and frequently also on conditions external to radiology. This must be demonstrated with specifically designed studies. As a matter of fact, we might have a fantastic test for the early diagnosis of disease X but, if no therapy exists for that disease, no impact on patient outcomes can be obtained. HTA should examine the link between each level and the next in the chain of this hierarchy to establish the clinical value of a radiologic test.

Cost-effectiveness in HTA

Cost-effectiveness should be included in HTA at any level of the hierarchic scale as cost per examination (level 1), per correct diagnosis (level 2), per invasive test avoided (level 3), per changed therapeutic plan (level 4), per gained

Table 0.1. Hierarchy of studies on diagnostic tests

Level	Parameters under investigation
6. Societal impact	Cost-benefit and cost-effectiveness analysis from a social perspective
5. Patient outcomes	Fraction of patients improved with the test compared with fraction improved without the test; difference in morbidity between the patients with the test and those without the test; gain in quality-adjusted life years (QALYs) obtained by the patients with the test compared with those without the test
4. Therapeutic impact	Fraction of patients for whom the test is judged useful for treatment planning or for whom the treatment planning is modified on the basis of the information supplied by the test
3. Diagnostic impact	Fraction of patients for whom the test is judged useful for reaching the diagnosis or for whom the diagnosis is substantially modified after the test; positive and negative likelihood ratios
2. Diagnostic performance	Sensitivity, specificity, accuracy, positive predictive value, negative predictive value and receiver operator characteristic (ROC) analysis; intraobserver, interobserver and interstudy reproducibility
1. Technical performance	Gray scale range; modulation transfer function change; sharpness; spatial resolution, in-plane (line pairs per mm, pixel size) and through-the-plane (slice thickness), integrated in voxel size; signal-to-noise ratio; contrast resolution (contrast-to-noise ratio); time resolution (images/sec) etc

Sources: Thornbury, 1994; Sunshine and Applegate, 2004; with modifications. In particular, reproducibility studies were added at level 2.

quality-adjusted life expectancy or per saved life (levels 5-6) [Hollingworth and Jarvik, 2007].

New equipment or a new imaging procedure should have extensive HTA assessment before it is adopted in routine practice. Thereafter a period of clinical evaluation follows where diagnostic accuracy is assessed against a known gold standard. Indeed, the radiologic literature is mainly composed of level 1 (*technical performance*) and level 2 (*diagnostic performance*) studies. This is partly inevitable. The evaluation of the technical and diagnostic performance of medical imaging is a typical function of radiologic research. However, radiologists less frequently study the *diagnostic impact* (level 3) or *therapeutic impact* (level 4) of medical imaging, while *outcome* (level 5) and *societal impact* (level 6) analysis is positively rare in radiologic research. A "shortage of coherent and consistent scientific evidence in the radiology literature" to be used for a wide application of EBR was noted in 2001 [Evidence-Based Radiology Working Group, 2001]. In recent years, several papers have appeared exploring levels higher than those concerning technical and diagnostic performance, such as the Scottish Low Back Pain Trial, the DAMASK study, and others [Gilbert FJ et al, 2004; Brealey et al, for the DAMASK Trial Team, 2007; Oei et al, 2008; Ouwendijk et al, 2008].

Shortage of scientific evidence in radiology

This lack of evidence on patient outcomes is a void also for well established technologies. This is the case for cranial CT for head injuries, even though the diagnostic information yielded by CT was "obviously so much better than [that] of alternative strategies that equipoise (genuine uncertainty about the efficacy of

a new medical technology) was never present" and "there was an effective treatment for patients with subdural or epidural hematomas – i.e. neurosurgical evacuation" [Hollingworth and Jarvik, 2007]. However, cases like this are very rare, and "in general, new imaging modalities and interventional procedures should be viewed with a degree of healthy skepticism to preserve equipoise until evidence dictates otherwise" [Hollingworth and Jarvik, 2007].

This urgent problem was recently highlighted by Christiane K. Kuhl and coworkers for the clinical value of 3.0-T MR imaging. They state: "Although for most neurologic and angiographic applications 3.0 T yields technical advantages compared to 1.5 T, the evidence regarding the added clinical value of high-field strength MR is very limited. There is no paucity of articles that focus on the technical evaluation of neurologic and angiographic applications at 3.0 T. This technology-driven science absorbs a lot of time and energy – energy that is not available for research on the actual clinical utility of high-field MR imaging" [Kuhl et al, 2008]. The same can be said for MR spectroscopy of brain tumors [Jordan et al, 2003; Hollingworth and Jarvik, 2007], with only one [Möller-Hartmann et al, 2002] of 96 reviewed articles evaluating the additional value of MR spectroscopy which compared this technology with MR imaging alone.

Reasons for shortage of high level radiologic studies

There are genuine reasons for rarely attaining the highest impact levels of efficacy by radiologic research. On the one hand, increasingly rapid technologic development forces an endless return to low impact levels. Radiology was judged as the most rapidly evolving specialty in medicine [Dixon, 1997]. On the other hand, level 5 and 6 studies entail long performance times, huge economic costs, a high degree of organization and management for longitudinal data gathering on patient outcomes, and often require a randomized study design (by way of example, the average time for 59 studies in radiation oncology up to publication of the results reviewed in 2005 was about 11 years [Soares et al, 2005]). In this setting, there are two essential needs: full cooperation with clinicians who manage the patient before and after a diagnostic examination, and methodologic and statistical expertise regarding randomized controlled trials. Radiologists should not be afraid of this, as it is not unfamiliar territory for radiology. More than three decades ago, mammographic screening created a scenario in which the early diagnosis by imaging contributed to a worldwide reduction in mortality from breast cancer, with a high societal impact.

Lastly, alternatives to clinical trials and meta-analyses exist. They are the so-called "pragmatic" or "quasi-experimental" studies and "decision analysis".

Pragmatic studies

A *pragmatic study* proposes the concurrent development, assessment, and implementation of new diagnostic technologies [Hunink and Krestin, 2002]. An empirically based study, preferably using controlled randomization, integrates research aims in clinical practice, using outcome measures reflecting the clinical decision-making process and acceptance of the new test. Outcome measures include: additional imaging studies requested; costs of diagnostic work-up and treatments; confidence in therapeutic decision-making; recruitment rate; and patient outcome measures. Importantly, time is used as a fundamental dimension, as an explanatory variable in data analysis to model the learning curve, technical developments, and interpretation skill. Limitations of this approach can be the need of dedicated and specifically trained personnel

and the related economic costs to be covered by presumably governmental agencies [Jarvik, 2002]. However, this seems to demonstrate the potential for responding to the dual demand of the increasing pace of technologic development in radiology and the need to attain higher levels of radiologic studies, thus in a single approach obtaining data on diagnostic confidence, effect on therapy planning, patient outcome measures and cost-effectiveness analysis.

Decision analysis

Decision analysis, based on deductive reasoning, tries to overcome the limited external validity associated with clinical trials [Hunink et al, 2001; Launois, 2003]. It is a tool for evaluating a diagnostic test on the basis of patient outcome using intermediate outcome measures such as sensitivity and specificity obtained by already published studies. Different diagnostic or therapeutic alternatives are visually represented by means of a decision tree and dedicated statistical methods are used (e.g. Markov model, Monte Carlo simulation) [Plevritis, 2005; Hunink et al, 2001]. This method is typically used for cost-effectiveness analysis. For instance, it was recently used for simulating the effectiveness of mammography, MR imaging, or both for screening of breast cancer in women carriers of BRCA1 mutations [Lee et al, 2008].

Looking at the pharmacologic research

A simple way to appraise the intrinsic difficulty in HTA of radiologic procedures is to compare radiologic with pharmacologic research (see Chapter 8). After the chemical discovery of an active molecule, its development, cell and animal testing, the phase I and phase II studies are carried out by the industry with the participation of very few clinicians (for phase I and II studies). Very few academic institutions and large hospitals are involved in this long phase (commonly about ten years). When clinicians become involved in phase III studies, i.e. large randomized trials for registration, the study aims have already reached level 5 (outcome impact). Radiologists have to climb 4 levels of impact before reaching the outcome level. Of course it is possible to imagine a world in which even radiologic procedures are tested for outcome endpoints before entering clinical practice, but the real world is different, such that we have much more technology-driven research from radiologists than radiologist-driven research on technology.

Why do we Need Biostatistics?

The application of EBM implies a fundamental difficulty. Not only producing scientific evidence but also reading and correctly understanding the medical literature, in particular summaries of the best results such as systematic reviews and meta-analyses, requires a basic knowledge of and confidence with the principles and techniques of *biostatistics*. In fact, this is the only way to quantify the uncertainty associated with biological variability and the changes brought about by the patient's disease. This theoretical background is now emerging as very important expertise to be acquired by any physician of the new millennium.

Descriptive statistics

Quantification of data variability and its presentation comprise the field of *descriptive statistics*. This branch of statistics enables us to describe the sample under investigation by summarizing its features with diagrams or graphs and various parameters (mean, standard deviation, median, etc.). The quantification of uncertainty is needed to understand the probability we have if we apply the results of a study to the general population from which the study sub-

Inferential statistics

jects were drawn, i.e. when we use *inferential statistics*. This allows us to propose a general view and a theoretical model of the phenomenon under investigation. In this way, we can anticipate future events, namely we can make an *inference*. This is a deduction which evaluates whether the results of a study on a sample size can be applied to the general population, with a controlled error probability. As a consequence, there is a close proximity between inferential statistics and *probability theory*.

Although biostatistics uses mathematical tools, which may be very simple or quite sophisticated, the problem is never a question of simple *mechanical computing* (today many software packages can adequately do the job). It is rather a question of understanding the meaning of the figures we obtain and the way we obtain them, both theoretically (what precisely do we mean by *specificity* or *likelihood ratio*?) and practically, for clinical decision-making.

Difference between statistical significance and clinical relevance

Note that while a statistically significant result can be lacking clinical relevance, clinically relevant evidence should be based on statistical significance. A study can produce very high statistical significance without having any clinical utility. Who would use an anti-hypertensive drug which systematically (i.e. in all subjects) reduces arterial pressure by 1 mmHg compared with standard treatment? On the other hand, the considerable effect of a new drug against a form of cancer, if real, will be demonstrated in a controlled study (i.e. compared with standard treatment) which shows a significant increase in the disease-free interval or survival time. In other words, the size of a statistically significant difference needs to be evaluated to conclude that it is also clinically relevant, while a clinically relevant difference, to become evidence, must produce a statistically significant difference in a high-quality study.

A particular aspect plays a role in clinical radiologic research. Even for simple studies on diagnostic performance (sensitivity, specificity etc.), the common lack of assumptions needed for applying parametric statistical methods (based on the direct computing of measured data) makes nonparametric statistical methods (based on qualitative classes or ranks, or other tools) frequently needed. However, understanding nonparametric statistics requires preliminary knowledge of basic parametric statistics.

There are several reasons for the prevalent use of nonparametric statistical methods in radiology. The most important are as follows: the frequent use of nominal scales of measurement, often simply dichotomous (positive or negative) or ordinal (a typical example is the Breast Imaging Reporting and Data System, BI-RADS®, scale [American College of Radiology, 2003]); the limited possibility of demonstrating the normal distribution of continuous numerical data in a small sample size (a necessary assumption for using parametric statistical methods); and the high frequency of a small sample size. As a consequence, most books concerning general medical statistics appear barely appropriate for radiologists. These texts commonly dedicate numerous pages to parametric methods and very few pages to nonparametric methods, and even when nonparametric methods are extensively explained, no specific reference to their use in diagnostic imaging is available. An exception to this trend in Italy is *Guida alla Statistica nelle Scienze Radiologiche* by Professor Guido Galli, a nuclear physician from the University of Rome School of Medicine [Galli, 2002].

The Structure of this Book

All these reasons explain the need for radiologists to possess knowledge in applied biostatistics. In the following chapters we will propose such knowledge, giving greater priority to logic than to computing.

Chapter 1: diagnostic performance

In *Chapter one* we describe the classic tools for the quantification of diagnostic performance typically used in radiologic studies: *sensitivity*, *specificity*, *predictive values*, *overall accuracy*, and *receiver operator characteristic* (ROC) *curve*. Moreover, we introduce here the *likelihood ratios* which quantify the *power* of a diagnostic test, i.e. the ability of the test to modify the disease (or non-disease) probability, up to now rarely used in radiologic studies. In this context, we show some aspects of probability theory and present Bayes' theorem.

Chapter 2: variables, scales of measurement, normal distribution, confidence intervals

In *Chapter two* we define the concept of *variable* and the different types of variables with reference to their *scales of measurement*, as well as some essential principles of *descriptive statistics*, *normal distribution*, and *confidence intervals*. Indeed, understanding the scales of measurement is essential for choosing a statistical test applicable to the data to be analyzed. Being familiar with normal distribution is a must for the use of all tools in biostatistics. Confidence intervals can be thought of as a conceptual and practical bridge between descriptive and inferential statistics: they define a range of variability in the results in the event the same study were to be repeated for a sample with the same size of patients having the same characteristics. An important trend in recent times is the increasing emphasis radiologic journals have been giving to confidence intervals. The presentation of the 95% confidence intervals should be considered mandatory for all indices of diagnostic performance.

Chapter 3: null hypothesis and statistical significance

Chapter three is dedicated to the theory of the scientific experiment, namely to the *null hypothesis* and *statistical significance*. This topic has the greatest philosophic and methodologic implications. We explain why the demonstration of an *experimental hypothesis* (e.g. that two diagnostic options have a different sensitivity for a given disease) must be obtained by working on an antithetical hypothesis (i.e. that there is no difference in sensitivity between the two diagnostic options) which we name null hypothesis. The researcher's aim is to demonstrate that the null hypothesis is sufficiently improbable to accept the indirect conclusion that the experimental hypothesis is probably true. However, this conclusion is never demonstrated in a direct and definitive way.

Chapter 4: parametric statistics

While in *Chapter four* we provide some of the essentials of parametric statistics and the assumptions required for the application of *parametric statistical tests*, in *Chapter five* we describe the most important *nonparametric statistical tests* and the assumptions needed for their application.

Chapter 5: nonparametric statistics

Chapter 6: association, correlation and regression

In *Chapter six* we define the concepts of *association*, *correlation*, and *regression* and propose the techniques for their quantification. Particular attention is paid to the differentiation between association or correlation between two variables and the deduction of the *cause-effect relationship*, the latter never being provable on the basis of a statistical calculation alone.

Chapter 7: reproducibility

In *Chapter seven* we present the most important techniques for evaluating the *reproducibility* of the result of a diagnostic test, either for continuous variables (Bland-Altman analysis), or for nominal and ordinal variables (Cohen k). Here

we introduce the concept of *intraobserver and interobserver variability*. These kinds of studies are currently highly appreciated for their ability to define the practical role of old and new imaging techniques.

Chapter 8: study design, sample size, systematic reviews (meta-analysis), levels of evidence

In *Chapter eight* the reader will find the principal *types of study* in relation to their *design* (observational or randomized experimental; prospective or retrospective; longitudinal or transversal; etc.) as well as a general description of the methods for calculating the *sample size*, i.e. the number of patients which need to be enrolled in a prospective study in order to obtain an acceptable probability of demonstrating the experimental hypothesis. Here we also include a short section on *systematic reviews*, namely those studies which gather together the information contained in already published studies on a given topic, conduct a critical appraisal of the methods used in those studies, select the studies according to predefined quality standards, and pool the results of the selected studies to provide a new and more reliable overall result, using dedicated statistical methods (*meta-analysis*). Afterwards, we define the so-called *levels of evidence* of radiologic studies.

Chapter 9: bias

In *Chapter nine* we present a list (without doubt incomplete) of the errors to be avoided in radiologic studies. In other words, we list the potential sources of *bias* that should be recognized as readers and avoided or, at least, limited and explicitly acknowledged as authors.

Chapter 10: recommendations for writing a radiologic paper

Finally, in *Chapter ten* we provide a series of practical recommendations for writing a radiologic study, with particular reference to the content of the four sections of the body of the paper and its logical structure (*Introduction*, *Materials and Methods*, *Results*, and *Discussion*) and the two essential accompanying items (*Abstract* and *References*).

What you will not find in this book

The subject matter of this book clearly falls short of exhaustively treating biostatistics in radiology, in part because radiology is transversally cross-linked with all medical subspecialties. However, a number of statistical techniques which can be used in radiologic research are not considered in this book. For example, the reader will not find suggestions for statistical and graphical methods for describing data; moreover, logistic regression, multiple regression, the concept of absolute and relative risk, survival curves, and non-inferiority studies are not treated. We have avoided these topics in order to produce a book capable of introducing biostatistics to radiologists. While it is only a preliminary approach to biostatistics, the volume does have the advantage of presenting the topic from the particular viewpoint of radiology.

Do not skip the examples

All the examples are progressively numbered in each chapter and drawn from the radiologic literature or invented ad hoc to facilitate the reader's understanding of the theoretical concepts. We recommend that the reader who has grasped a theoretical definition should not skip the examples, as they could be a useful aid for committing the theoretical concept to memory. Similarly, we advise the reader who is having difficulty coming to grips with the theoretical definition to go straight to the following example as this could immediately shed light on the theoretical problem.

Formulas

A final word of advice. Throughout the book the reader will find several mathematical formulas. These have been included in their entirety for the readers willing to understand the mechanism of computing. However, a thorough understanding of the formulas is by no means required to grasp the general sense of the concepts and their practical use.

Radiologists interacting with statisticians

It is far from our intention to educate radiologists so that they can replace statisticians, as this appears neither possible nor useful. Instead it is our aim educate radiologists so that they may interact with statisticians with proficiency and critical judgment.

References

Acheson L, Mitchell L (1993) The routine antenatal diagnostic imaging with ultrasound study. The challenge to practice evidence-based obstetrics. Arch Fam Med 2:1229-1231

American College of Radiology (2003) ACR breast imaging reporting and data system (BI-RADS): Breast Imaging Atlas. American College of Radiology, Reston, Va

Arrivé L, Tubiana JM (2002) "Evidence-based" radiology. J Radiol 83:661

Barr HJ, Ohlhaber T, Finder C (2006) Focusing in on dose reduction: the FDA perspective. AJR Am J Roentgenol 186:1716-1717

Bennett JD (2003) Evidence-based radiology problems. Covered stent treatment of an axillary artery pseudoaneurysm: June 2003-June 2004. Can Assoc Radiol J 54:140-143

Blackmore CC (2003) Evidence-based imaging evaluation of the cervical spine in trauma. Neuroimaging Clin N Am 13:283-291

Blackmore CC (2004) Critically assessing the radiology literature. Acad Radiol 11:134-140

Brealey SD; DAMASK (Direct Access to Magnetic Resonance Imaging: Assessment for Suspect Knees) Trial Team (2007) Influence of magnetic resonance of the knee on GPs' decisions: a randomised trial. Br J Gen Pract 57:622-629

Bui AA, Taira RK, Dionisio JD et al (2002) Evidence-based radiology: requirements for electronic access. Acad Radiol 9:662-669

Centre for Evidence-Based Medicine, Oxford University, England. http://cebm.net

Cohen WA, Giauque AP, Hallam DK et al (2003) Evidence-based approach to use of MR imaging in acute spinal trauma. Eur J Radiol 48:49-60

Council of the European Union (1997) Council Directive 97/43/Euratom of 30 June 1997 on health protection of individuals against the dangers of ionizing radiation in relation with medical exposure, and repealing Directive 84/466/Euratom. J Eur Commun L 180:22-27 (http://europa.eu.int/eur-lex/en/dat/1997/en_397L0043.htlm)

Dixon AK (1997) Evidence-based diagnostic radiology. Lancet 350:509-512

Dodd JD (2007) Evidence-based practice in radiology: steps 3 and 4 – Appraise and apply diagnostic radiology literature. Radiology 242:342-354

Dodd JD, MacEneaney PM, Malone DE (2004) Evidence-based radiology: how to quickly assess the validity and strength of publications in the diagnostic radiology literature. Eur Radiol 14:915-922

Erden A (2004) Evidence based radiology Tani Girisim Radyol 10:89-91

Evidence-Based Radiology Working Group (2001) Evidence-based radiology: a new approach to the practice of radiology. Radiology 220:566-575

FDA Radiological Health Program (2008). Available at: http://www.fda.gov/cdrh/radhealth/index.html

Fineberg HV, Bauman R, Sosman M (1977) Computerized cranial tomography. Effect on diagnostic and therapeutic plans. JAMA 238:224-227

Fryback DG, Thornbury JR (1991) The efficacy of diagnostic imaging. Med Decis Making 11:88-94

Galli G (2002) Guida alla statistica nelle scienze radiologiche. Ecoedizioni internazionali, Rome, Italy

Gilbert FJ, Grant AM, Gillan MGC (2004) Low back pain: influence of early MR imaging or CT on treatment and outcome – multicenter randomized trial. Radiology 231:343-351

Giovagnoni A, Ottaviani L, Mensà A et al (2005) Evidence based medicine (EBM) and evidence based radiology (EBR) in the follow-up of the patients after surgery for lung and colon-rectal carcinoma. Radiol Med 109:345-357

Goergen SK, Fong C, Dalziel K, Fennessy G (2003) Development of an evidence-based guideline for imaging in cervical spine trauma. Australas Radiol 47:240-246

Greenhalgh T (2006) How to read a paper. The basics of evidence-based medicine. 3rd ed. Blackwell, Oxford, England: ix-xii (a); 1-3 (b)

Guillerman RP, Brody AS, Kraus SJ (2002) Evidence-based guidelines for pediatric imaging: the example of the child with possible appendicitis. Pediatr Ann 31:629-640

Hillman BJ, Gatsonis CA (2008) When is the right time to conduct a clinical trial of a diagnostic imaging technology? Radiology 248:12-15

Hollingworth W, Jarvik JG (2007) Technology assessment in radiology: putting the evidence in evidence-based radiology. Radiology 244:31-38

Hunink MG, Glasziou PP, Siegel JE et al (2001) Decision making in health and medicine: integrating evidence and values. Cambridge University Press, Cambridge, UK, 2001

Hunink MG, Krestin GP (2002) Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. Radiology 222:604-614

Jarvik JG (2002) Study design for the new millennium: changing how we perform research and practice medicine. Radiology 222:593-594

Jordan HS, Bert RB, Chew P et al (2003) Magnetic resonance spectroscopy for brain tumors. Agency for Healthcare Research and Quality, Rockville, MD:109

Kainberger F, Czembirek H, Frühwald F et al (2002). Guidelines and algorithms: strategies for standardization of referral criteria in diagnostic radiology. Eur Radiol 12:673-679

Kuhl CK, Träber F, Schild HH (2008) Whole-body high-field-strength (3.0-T) MR imaging in clinical practice. Part I. Technical considerations and clinical applications. Radiology 246:675-696

Launois R (2003) Economic assessment, a field between clinical research and observational studies. Bull Cancer 90:97-104

Lee JM, Kopans DB, McMahon PM et al (2008). Breast cancer screening in BRCA1 mutation carriers: effectiveness of MR imaging – Markov Monte Carlo decision analysis. Radiology 246:763-771

Liedberg J, Panmekiate S, Petersson A, Rohlin M (1996) Evidence-based evaluation of three imaging methods for the temporomandibular disc. Dentomaxillofac Radiol 25:234-241

Mackenzie R, Dixon AK (1995) Measuring the effects of imaging: an evaluative framework. Clin Radiol 50:513-518

Malone DE (2007) Evidence-based practice in radiology: an introduction to the series. Radiology 242:12-14

Malone DE, Staunton M (2007) Evidence-based practice in radiology: step 5 (evaluate) – Caveats and common questions. Radiology 243:319-328

Matowe L, Gilbert FJ (2004) How to synthesize evidence for imaging guidelines. Clin Radiol 59:63-68

Medina LS, Aguirre E, Zurakowski D (2003) Introduction to evidence-based imaging. Neuroimaging Clin N Am 13:157-165

Medina LS, Blackmore CC (2006) Evidence-based imaging. 1st edn. Springer, New York, NY

Medina LS, Blackmore CC (2007) Evidence-based radiology: review and dissemination. Radiology 244:331-336

Möller-Hartmann W, Herminghaus S, Krings T et al (2002) Clinical application of proton magnetic resonance spectroscopy in the diagnosis of intracranial mass lesions. Neuroradiology 44:371-381

Mukerjee A (1999) Towards evidence based emergency medicine: best BETs from the Manchester Royal Infirmary. Magnetic resonance imaging in acute knee haemarthrosis. J Accid Emerg Med 16:216-217

No authors listed (1997) Reports from the British Columbia Office of Health Technology Assessment (BCOHTA). Routine ultrasound imaging in pregnancy: how evidence-based are the guidelines? Int J Technol Assess Health Care 13:633-637

No authors listed (1997) Routine ultrasound imaging in pregnancy: how evidence-based are the guidelines? Int J Technol Assess Health Care 13:475-477

No authors listed (2004) Proceedings of the Second ALARA Conference. February 28, 2004. Houston, Texas, USA. Pediatr Radiol 34[Suppl 3]:S162-246

Oei EH, Nikken JJ, Ginai AZ et al; From the Program for the Assessment of Radiological Technology (ART Program) (2008) Costs and effectiveness of a brief MRI examination of patients with acute knee injury. Eur Radiol 2008 Sep 16. [Epub ahead of print]

Ouwendijk R, de Vries M, Stijnen T et al; From the Program for the Assessment of Radiological Technology (2008) Multicenter randomized controlled trial of the costs and effects of noninvasive diagnostic imaging in patients with peripheral arterial disease: the DIPAD trial. AJR Am J Roentgenol 190:1349-1357

Plevritis SK (2005) Decision analysis and simulation modeling for evaluating diagnostic tests on the basis of patient outcomes. AJR Am J Roentgenol 185:581-590

Prasad KN, Cole WC, Haase GM (2004) Radiation protection in humans: extending the concept of as low as reasonably achievable (ALARA) from dose to biological damage. Br J Radiol 77:97-99

Raymond J, Trop I (2007) The practice of ethics in the era of evidence-based radiology. Radiology 244:643-649

RCR Working Party (1998) Making the best use of a department of clinical radiology: guidelines for doctors. 4th edn. The Royal College of Radiologists, London

Sackett DL, Rosenberg WM, Gray JA et al (1996) Evidence based medicine: what it is and what it isn't. BMJ 312:71-72

Semelka RC, Armao DM, Elias J Jr, Huda W (2007) Imaging strategies to reduce the risk of radiation in CT studies, including selective substitution with MRI. J Magn Reson Imaging 25:900-909

Soares HP, Kumar A, Daniels S et al (2005) Evaluation of new treatments in radiation oncology: are they better than standard treatments? JAMA 293:970-978

Sunshine JH, Applegate KE (2004) Technology assessment for radiologists. Radiology 230:309-314

Taïeb S, Vennin P (2001) Evidence-based medicine: towards evidence-based radiology. J Radiol 82:887-890

Thornbury JR (1994) Clinical efficacy of diagnostic imaging: love it or leave it. AJR Am J Roentgenol 162:1-8

Thornbury JR (1999) Intermediate outcomes: diagnostic and therapeutic impact. Acad Radiol 6[suppl 1]:S58-S65

Tonelli MR (1998) The philosophical limits of evidence-based medicine. Acad Med 73:1234-1240

Trinder L (2000) A critical appraisal of evidence-based practice. In: Trinder L, Reynolds S (eds) Evidence-based practice: a critical appraisal. Blackwell Science, Oxford, England:212-214

van Beek EJ, Malone DE (2007) Evidence-based practice in radiology education: why and how should we teach it? Radiology 243:633-640

White SJ, Ashby D, Brown PJ (2000) An introduction to statistical methods for health technology assessment. Health Technol Assess 4:i-iv, 1-59

Woolf SH, Grol R, Hutchinson A et al (1999) Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. BMJ 318:527-530

1

Diagnostic Performance

Don't stop with an answer [...]
An answer is always the stretch of road that's behind you.
Only a question can point the way forward.

JOSTEIN GAARDER

Diagnostic performance

Measures of diagnostic performance

The *performance* of a diagnostic examination[1] can be basically considered as its degree of *accuracy*, namely its ability to find the subjects affected with a given disease as positive and the subjects not affected with same disease as negative. The indices which in different ways measure this performance are defined *measures of diagnostic performance* and the studies aimed at measuring the diagnostic performance of an examination or, more often, at comparing the diagnostic performance of two or more examinations, are defined *studies of diagnostic performance*.

Firstly we will present the five most commonly used indices of diagnostic performance in radiologic papers: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and overall accuracy. Thereafter, we will consider *receiver operator characteristic* (ROC) curves, which are also widely used in radiologic papers, and *likelihood ratios*, which are special indices which quantify the ability of a diagnostic examination to change the disease probability (i.e. the *power* of the diagnostic examination) and which to date have been little used in radiologic papers. In this setting we will show some features of probability theory and Bayes' theorem. For the sake of clarity, the likelihood ratio will be explained before ROC analysis.

[1] For the sake of clarity, we will avoid naming a radiologic examination as a *test* as much as possible. Although this term is entirely correct, we prefer to use the term *exam* or *examination* in order to avoid confusion with *statistical tests*. Exceptions will be *pretest probability* and *post-test probability* for a given disease which we will approach with Bayes' theorem to represent the ability of each diagnostic test to increase or decrease the probability of a given disease in the subjects who underwent the examination with a positive or a negative result, respectively. Other rare exceptions will be evident from the context.

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

1.1. The Results of an Examination Compared to a Reference Standard

Reference standard

If we want to evaluate the performance of a diagnostic examination, we need to compare its results to a *reference standard*, a term which today is preferable to *gold standard*, since the latter is considered exceedingly optimistic (as in other fields, in Biostatistics *all that glistens is not golden*). In oncologic diagnostics, the typical example is to verify each result of a diagnostic examination for a sample of *n* patients with the pathology report, both of which refer to a defined lesion. Suppose that both the radiologist and the pathologist are required to give a *dichotomous* judgment (yes/no) about the malignancy of a lesion. In this case, the pathology examination is the reference standard and states whether each result of the diagnostic examination is true or false. It will be *true positive* when the radiologist has correctly defined a pathologically proven malignant lesion as positive, *true negative* when the radiologist has correctly defined a non-malignant finding as negative, *false positive* when the radiologist has incorrectly defined a non-malignant finding as positive, and *false negative* when the radiologist has incorrectly defined a malignant lesion as negative. The *n* cases which make up the sample of this comparison are distributed among these four possibilities according to the rule that each case is assigned to only one of the four categories. Using these data, we can generate a *two-by-two contingency table* where the number of true positives, false positives, false negatives, and true negatives are reported (Table 1.1).

True positive, true negative, false positive, false negative

Two-by-two contingency table

Note that this table can be completed with the total of the lines and with the total of the columns, namely with a series of marginal totals (all the positive cases at the diagnostic examination; all the negative cases at the diagnostic examination; all the positive cases at the reference standard; and all the negative cases at the reference standard), and with the grand total of the *n* patients or subjects under investigation, as shown in Table 1.2.

Cases, lesions, findings, patients, subjects

The careful reader has probably realized that we have intermingled different terms: *cases*, *lesions*, *findings*, *patients*, and *subjects*. Pay attention to the meaning of these words. In a scientific paper, these terms cannot be interchanged and one of them (*cases*) should be carefully avoided. We can properly consider the study subjects as *patients* when they present with symptoms or signs for a disease. On the other hand, we name the asymptomatic persons enrolled in population screening program only as *subjects*. However, it is cor-

Table 1.1. Two-by-two contingency table for the comparison between the results of a radiologic examination and those of a reference standard

		Reference standard	
		Positive	**Negative**
Radiologic examination	**Positive**	True positives (TP)	False positives (FP)
	Negative	False negatives (FN)	True negatives (TN)

Table 1.2. Two-by-two contingency table for the comparison between the results of a radiologic examination and those of a reference standard in a series of subjects, completed with marginal totals and grand totals

		Reference standard		
		Affected	**Nonaffected**	**Total**
Radiologic examination	**Positive**	True positives (TP)	False positives (FP)	All positives (TP + FP)
	Negative	False negatives (FN)	True negatives (TN)	All negatives (FN + TN)
Total		All affected (TP + FN)	All nonaffected (FP + TN)	Grand total (TP + FP + FN + TN)

rect to name a group of patients as subjects. Words mean things: the frequency of disease is certainly greater in patients than in symptomatic subjects, with relevant practical consequences which we will see. However, the distinction between patients and subjects is relatively trivial.

The statistical unit to be measured

More importantly, we should fully understand what changes when the *statistical unit* is no longer the patient (or the subject) but each lesion (or finding). Of course, if each patient has no lesions or only one lesion, we have no consequences in statistical calculations. But a patient can have more than one lesion, as typically we find in the study of liver metastases. The same reasoning can be applied to each of the two kidneys, breasts, lungs, or to a single lobe or segment of the brain, liver, lung, prostate, coronary tree, etc. We should always be extremely clear regarding the application of the indices of diagnostic performance. On what basis are they calculated? Patient by patient? Organ by organ? Segment by segment? Lesion by lesion? Note that the term *case* is ambiguous because it can be used for both patients and lesions. It should therefore be strictly avoided in a scientific context. Refer your description to the real statistical units under investigation.

Avoid the term *case* in a scientific context

Initial studies versus large clinical studies

We can at this point note the value of a general principle. The initial studies on the diagnostic performance of a new imaging modality or technique benefit greatly from the reporting of indices on a lesion-by-lesion basis: we can have a small number of patients and obtain a measure of what happens for each of the lesions. Afterwards, more conclusive information on the value of the clinical application of a new modality or technique can be obtained with a patient-by-patient analysis. In the latter situation, we sometimes have to solve relevant conceptual problems (implying the clinical relevance of radiologic findings) for the definition of true positive, false positive, true negative, and false negative patient when multiple lesions are present and lobes, segments or organs are affected by the disease.

1.2. Measures of Diagnostic Performance

Using the figures of the true positives, false positives, true negatives, and false negatives, we can calculate a series of indices which measure the diagnostic performance. Table 1.3 reports definitions and formulas of these indices, as well as their dependence or independence on disease prevalence.

Table 1.3. Indices measuring diagnostic performance

Index	Definition	Formula	Dependence on disease prevalence
1. Sensitivity (or TP rate)	Ability to identify the presence of disease	TP/(TP+FN)	No
2. Specificity (or TN rate)	Ability to identify the absence of disease	TN/(TN+FP)	No
3. Positive predictive value (PPV)	Reliability of the positive result	TP/(TP+FP)	Yes
4. Negative predictive value (NPV)	Reliability of the negative result	TN/(TN+FN)	Yes
5. Overall accuracy	Global reliability	(TP+TN)/(TP+TN+FP+FN)	Yes
6. FN rate	Proportion between FN and all affected	FN/(FN+TP) = (1 – Sensitivity)	No
7. FP rate	Proportion between FP and all nonaffected	FP/(FP+TN) = (1 – Specificity)	No
8. Positive likelihood ratio	Increase in disease probability when the result is positive	Sensitivity/(1 – Specificity)	No
9. Negative likelihood ratio	Decrease in disease probability when the result is negative	(1 – Sensitivity)/Specificity	No

Note that *disease prevalence* is equal to (TP+FN) / (TP+TN+FP+FN), being the ratio between the number of subjects affected by the disease and the grand total of sample of subjects under investigation.

All of these are simple proportions or ratios which differently combine the four quantities of the two-by-two contingency table. The first seven indices range between 0 and 1 and frequently are reported as percentages. The first five indices indicate an increasingly high diagnostic performance of the examination under investigation the closer they are to 1. The sixth and seventh indices indicate an increasingly high diagnostic performance of the examination under investigation the closer they are to 0. Moreover, they are frequently defined as *1 – sensitivity* and *1 – specificity*, respectively. The meaning of the last two indices, i.e. the likelihood ratios (LRs), is a bit more complex. They theoretically range between 0 and infinity but practically indicate an increasingly high diagnostic performance the further they move away from 1, with the positive LR moving towards values higher than 1 and the negative LR towards values lower than 1.

1.3. Sensitivity, Specificity, FN Rate and FP Rate

Sensitivity: the ability to identify the presence of a disease

The meaning of *sensitivity* is intuitive: *it is the ability of an examination to identify the presence of a given disease.* It can also be considered as the proportion between the number of positive subjects with the disease and the total

number of subjects with the disease, namely the proportion of subjects with the disease who were correctly detected by the radiologist. Sensitivity is given by the ratio TP/(TP + FN), i.e. the *proportion of positives among the subjects with the disease*.

If the number of true positives is unchanged, sensitivity is inversely related to the number of false negatives. In fact, the *false negative rate*, namely the proportion of subjects falsely considered nonaffected by the disease, summed with the sensitivity gives a result equal to 1. In other words, the false negative rate is the complement to 1 of sensitivity.

False negative rate

Example 1.1. Sensitivity of mammography and dynamic contrast enhanced magnetic resonance (MR) imaging for the detection of malignant lesions in patients candidate for mastectomy. The authors investigate 99 breasts in 90 candidates for unilateral (n = 81) or bilateral (n = 9) mastectomy. The reference standard, i.e. the pathology examination of the whole excised breast, establishes the presence of 188 malignant lesions. Mammography has 124 true positives and 64 false negatives, MR imaging 152 true positives and 36 false negatives. As a consequence, sensitivity is 124/(124+64) = 0.660 for mammography and 152/(152+36) = 0.809 for MR imaging. The lesion-by-lesion sensitivity of mammography is 66.0%, that of MR imaging is 80.9%. The FN rate is 0.340 or 34.0% and 0.191 or 19.1%, respectively. Note that the statistical unit is the lesion and not the patient or the breast [Sardanelli et al, 2004].

Specificity: the ability to identify the absence of a disease

False positive rate

The meaning of *specificity* is evident less immediately. It refers to *the ability of the examination to identify the absence of a given disease*, given by the ratio TN/(TN + FP), i.e. the proportion of the negatives among the subjects not affected with the disease. If the number of true negatives is unchanged, it is inversely related to the number of false positives. In fact, the *false positive rate*, i.e. the proportion of subjects falsely considered to be affected by the disease, summed with specificity gives 1. In other words, the false positive rate is the complement to 1 of specificity.

"Specificity" in common language

The less immediate understanding of the term specificity is due to its common improper use, at least in spoken language, to indicate the ability of an examination to make a certain diagnosis. This improper use often implies several logical mistakes. For instance, if we state that computed tomography (CT) is highly "specific" for the diagnosis of intracranial hemorrhage, we would mean that this imaging modality can reliably identify a hyperattenuation on nonenhanced scans as a hemorrhage. However, this statement has two different meanings: if really there is an intracranial hemorrhage, it is highly probable that CT can detect it; a CT diagnosis of intracranial hemorrhage is rarely a false positive.

Using correct scientific terminology, these two sentences are the same as saying that CT has both high sensitivity and high positive predictive value for intracranial hemorrhage. As both specificity and positive predictive value are inversely related to false positives, if we have very few false positives, it will be true that the examination will also be highly specific (under the condition of having a suitable number of true negatives). At any rate, *we cannot say that an examination is highly specific thinking that our audience also understands it to*

be highly sensitive. This is a conceptual error. If both sensitivity and specificity are high, the examination is highly accurate (not only highly specific), as CT actually is for intracranial hemorrhage.

Care must be taken, since an examination could have few false positives and many false negatives, thus at the same time being highly specific but not very sensitive. As a consequence, it will be of little use as a diagnostic tool in symptomatic patients, despite being highly specific.

Moreover, if we say CT is highly specific for the differentiation between acute intracranial hemorrhage and acute brain ischemia, we fall into a deeper complication. This sentence should imply a high sensitivity for both conditions, probably relatively higher for the former than for the latter due to the false negatives associated with tiny ischemias. Similarly, the specificity of CT will be different due to the relatively large number of hypoattenuations caused by previous infarcts in elderly patients or due to artifacts, etc, when compared with the hyperattenuations which can be falsely attributed to hemorrhage. The key point is that high CT specificity for acute intracranial hemorrhage does not imply high specificity for acute brain ischemia. The same reasoning can be applied to sensitivity. *For the sake of clarity, we should always distinguish between CT sensitivity and CT specificity for each of the two conditions.*

Example 1.2. Specificity. Low-dose CT screening for lung cancer. Of a total of 1611 asymptomatic subjects who undergo the first screening event, 186 are found to be positive and are further studied with high-resolution scanning; 21 of these undergo biopsy. Thirteen subjects are found to be affected by lung cancer. There are no interval cancers (cancers detected between the first and the second screening event). As a result there are 1425 true negatives (the total of 1611 minus 186 positives) and 173 false positives (186 positives minus 13 true positives). Specificity is 1425/(1425+173) = 1425/1598 = 0.892 = 89.2% [Sobue et al, 2002]. In this series only one possible lesion is considered for each subject. Lesion and subject are coincident as a statistical unit.

Sensitivity and specificity: answers to pretest questions

Sensitivity and specificity answer questions which can be raised before requesting or performing an examination. They therefore provide answers to aprioristic[2] questions:

- If the patient is affected by the disease, what is the probability that the examination produces a positive result (sensitivity)?
- If the patient is not affected by the disease, what is the probability that the examination produces a negative result (specificity)?

[2] The differentiation between *sensitivity and specificity* as *answers to pre-examination questions* on the one hand and *predictive values* as *answers to post-examination questions* on the other hand underlines the different logic of these indices of diagnostic performance. Sensitivity and specificity should be used to refer to the intrinsic diagnostic performance of a given examination while predictive values enable us to evaluate the reliability of the results of the same examination once it is permormed. It should be borne in mind that these are not the same thing, as we will explain by demonstrating the influence of disease prevalence on predictive values. Notice that *another pre-/post-examination differentiation* is related to the concepts of *pretest probability* and *post-test probability* which we will introduce for the application of Bayes' theorem (see Section 1.5).

Sensitivity and specificity do not depend on disease prevalence

Sensitivity and specificity (as well as the false negative rate and the false positive rate) depend on the technical characteristics of the examination, on the capability of the radiologist and her/his team (radiographers, nurses, etc.) to perform the examination, and on the radiologist's skill in interpreting the examination. Sensitivity and specificity are not influenced by the disease prevalence in the study population (they are instead influenced by the degree, the stage of the disease, as we will demonstrate in the next section). The term *prevalence* indicates the proportion between the number of subjects affected by a disease and the total number of subjects of an entire population (or of a sample, frequently named *study population*) for a defined time interval, whereas the term *incidence* indicates the number of subjects newly diagnosed as affected by the disease during a defined time interval (see the Note to Table 1.3).

Prevalence

Incidence

As a matter of fact, the optimal situation in clinical practice is when a single diagnostic examination is available with levels of sensitivity or specificity high enough to produce conclusive decision-making. These two extreme conditions are defined as follows: an examination is *SNOUT* when its negative result excludes the possibility of the presence of the disease (*when a test has a very high Sensitivity, a Negative result rules OUT the diagnosis*); it is instead *SPIN* when its positive result definitely confirms the presence of the disease (*when a test has a very high SPecificity, a positive result rules IN the diagnosis*). In most situations, a certain degree of certainty can be reached with a single diagnostic examination but not a definitive conclusion. More than one examination is generally needed. They are ordered according to a flow-chart which takes into account sex, age, familial and personal history, clinical history, previous examinations, etc.

SNOUT and SPIN

In other words, sensitivity and specificity alone cannot translate the result of a radiologic examination into clinical practice.

1.4. Predictive Values, Diagnostic Accuracy and Disease Prevalence

Predictive values: answers to post-test questions

A first possibility for the translation of the result of an examination in clinical practice is provided by *predictive values*. These indicate *the reliability of the positive or negative result* and answer questions posed *after* having performed the examination:

- If the result of the examination is positive, what is the probability that the patient really is affected by the disease (positive predictive value)?
- If the result of the examination is negative, what is the probability that the patient is really not affected by the disease (negative predictive value)?

Predictive values depend on disease prevalence

The predictive values depend not only on technical parameters and on the ability to perform the examination and interpret the results. In fact, if sensitivity and specificity are kept unchanged, predictive values change in relation with disease prevalence: the positive predictive value is directly related to disease prevalence whereas the negative predictive value is inversely related to disease prevalence.

Predictive values depend on disease prevalence. This is not intuitive and implies important practical consequences. Let us reflect upon this statement for a moment: when the disease prevalence is very low, a very high sensitivity is associated with a very low positive predictive value.

A useful way of envisaging this situation is provided by the following example. If all the sample subjects have the disease, the positive predictive value is always 1.0 (i.e. 100%) even with very low sensitivity (but not 0) and the negative predictive value is always 0.0 (i.e. 0%) even with very high specificity (even equal to 1.0, i.e. 100%). Similarly, if all the sample subjects do not have the disease, the negative predictive value is always 1.0 (i.e. 100%) even with very low sensitivity (but not 0) and the positive predictive value is always 0.0 (i.e. 0%) even with very high specificity (even equal to 1.0, i.e. 100%). It therefore follows that an examination with the highest possible sensitivity cannot correctly diagnose a non-existent disease, and an examination with the highest possible specificity cannot correctly diagnose the absence of disease in subjects who have the disease.

The reliability of our reports also depends on patient selection by the referring physicians

We can obtain increasingly higher levels of sensitivity and specificity, but the reliability of our reports (i.e. our predictive values) will depend on disease prevalence, namely on the epidemiologic context and, in clinical practice, on patient selection by the referring physician with a diagnostic query.

Now, we should at this point introduce a new variable to the system. *A disease can affect a patient with different levels of severity (or stage) and the probability of a positive result of an examination increases with the level of severity.* The level of severity should be lower in subjects in whom the disease is diagnosed with periodic screening than that found in symptomatic subjects in whom the disease is diagnosed in clinical practice. *In this way we observe a direct influence on sensitivity and specificity: they are higher in symptomatic subjects than in asymptomatic subjects in whom the disease is more likely in an early stage.* This difference is lower at the first round of an oncologic screening program (when we detect the *prevalent tumors*, with numerous cases which could have been diagnosed even in an earlier stage) and is higher in the later rounds (when we detect the *incident tumors*, not present at the first round). *Basically, subject selection, which determines the level of severity of the disease, also influences sensitivity and specificity.* We will return to this feature after introducing the concept of diagnostic threshold.

Overall accuracy: the ability to correctly identify the presence and the absence of a disease

Overall accuracy is the ability of an examination to correctly diagnose both subjects affected with the disease and subjects not affected with the disease as a fraction of the total number of examined subjects. It answers the question: what is the probability of a correct result? It is somewhat like a *global index of diagnostic performance*, but its linear distribution ranges between the sensitivity value and the specificity value. It approaches the higher of the two with increasing disease prevalence and approaches the lower of the two with decreasing disease prevalence. *In practice, it is a kind of "mean" between sensitivity and specificity which is weighted for disease prevalence.* Dependence on disease prevalence is the feature shared with the predictive values. The graphs in Figure 1.1. show the dependence of predictive values and overall accuracy on disease prevalence.

Example 1.3. Predictive values of clinical and screening mammography. Imagine 10,000 women with a palpable lump are studied (*clinical mammography*), with 95% sensitivity and 80% specificity. With a disease prevalence of 50%, we would have 4,750 true positives, 4,000 true negatives, 1,000 false positives, and 250 false negatives. The PPV would be 4,750/(4,750+1,000) = 0.826 = 82.6%; the NPV 4,000/(4,000+250) = 0.941 = 94.1%. For nearly every 5 women affected with cancer there would be a healthy woman who undergoes diagnostic work-up with possible needle

biopsy (4,750/1000 = 4.75). This woman with a benign palpable lump is unlikely to consider invasive examinations as useless or dangerous. However, if we were to study 10,000 asymptomatic women (*screening mammography*) with the same levels of sensitivity and specificity (95% and 80%, respectively) with a disease prevalence of 3%, we would have 285 true positives, 7,760 true negatives, 1,940 false positives, and 15 false negatives. The NPV would go up to 7,760/(7,760+15) = 0.998 = 99.8%, PPV would go down to 285/(285+1,940) = 0.128 = 12.8%. This means that nearly 7 healthy women would be sent for diagnostic work-up with a possible needle biopsy for every woman effectively diagnosed with cancer (1,940/285 = 6.8). The recall rate would be very high, equivalent to 22.25% (2,225/10,000). The overall effect would be a *false alarm* (if at every round we recall 20-25% of the women, after 4-5 rounds on average all the women would be recalled).

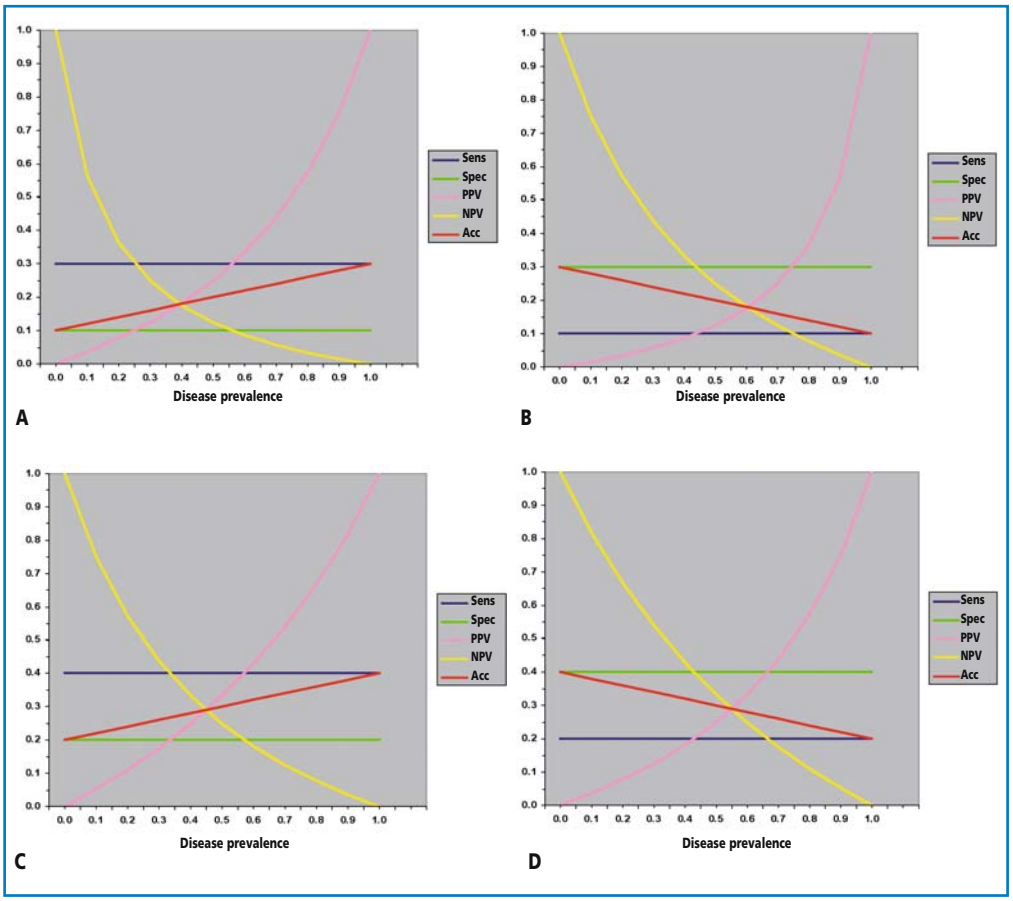
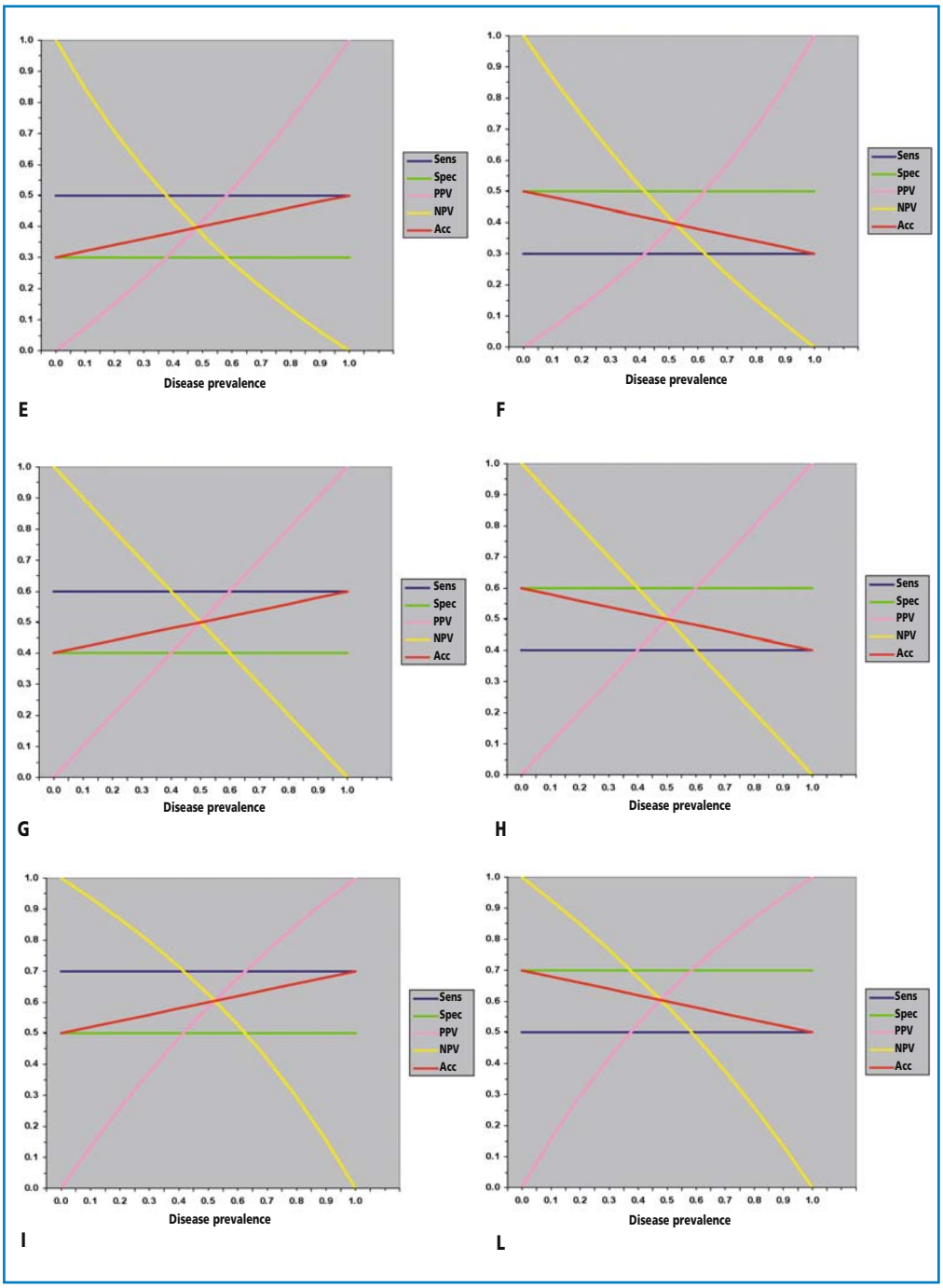


Figure 1.1. Distribution of positive predictive value (PPV), negative predictive value (NPV) and overall accuracy as a function of disease prevalence. The figure shows a series of paired graphs where the values of sensitivity and specificity are constant and represented by a blue and a green line, respectively. For the sake of clarity, the absolute difference between sensitivity and specificity is also constant, equal to 0.2. We present the following pairs of sensitivity and specificity values, respectively: 0.3, 0.1 (**A**) and vice versa 0.1, 0.3 (panel **B**); and so on, 0.4, 0.2 (**C**) and 0.2, 0.4 (**D**); 0.5, 0.3 (**E**) and 0.3, 0.5 (**F**); 0.6, 0.4 (**G**) and 0.4, 0.6 (**H**); 0.7, 0.5 (**I**) and 0.5, 0.7 (**L**); 0.8, 0.6 (**M**) and 0.6, 0.8 (**N**); 0.9, 0.7 (**O**) and 0.7, 0.9 (**P**).

(continued)

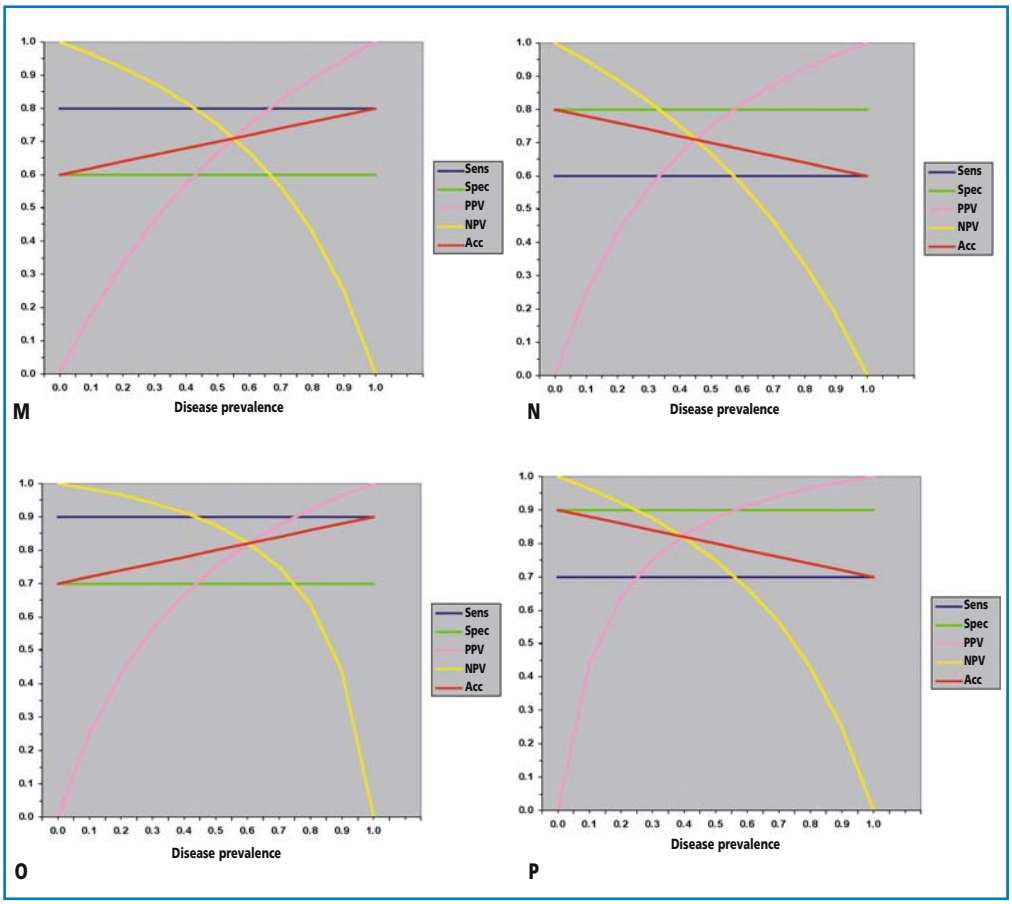
(continued)



Note that: (i) the behaviour of the overall accuracy curve is linear between the sensitivity and specificity values, ascending when sensitivity is higher than specificity (graphs on the left), descending when vice versa (graphs on the right); (ii) regardless of how high or low sensitivity and specificity are, PPV (yellow line) and NPV (pink line) always range between 0 and 1, with linear behaviour of the curve only in the particular case where sensitivity and specificity are equidistant from the horizontal middle line at 0.5 (**G**, **H**); (iii) PPV, NPV and overall accuracy intersect at 0.5 of disease prevalence when sensitivity and specificity are equidistant from the 0.5-horizontal midline (**G**, **H**). In this case also PPV, NPV and overall accuracy are

(continued)

(continued)



equal to 0.5 and the intersection is at the center of the graph (when sensitivity is equal to specificity, the intersection between PPV, NPV, and accuracy is at a disease prevalence of 0.5 but at a y-coordinate corresponding to the value of sensitivity and specificity, cases not shown); iv) for values of sensitivity and specificity near to those in clinical practice (i.e. over 0.5), PPV, NPV, and accuracy intersect with a prevalence higher than 0.5 (i.e. on the right) when sensitivity is higher than specificity, while it is with a prevalence lower than 0.5 (i.e. on the left) with sensitivity lower than specificity. With reference to panels **G** and **H**, we can also note that: i) likelihood ratios (see Section 1.5), which depend on sensitivity and specificity, are constant and equal to 1.0, i.e. the examination has no *power* but the disease prevalence generates a range between 0.0 and 1.0 for PPV and NPV, with linear curve behaviour; ii) if we were to progressively increase the difference between sensitivity and specificity, the slope of the red line (overall) would tend to overlay that of the two predictive values. The extreme cases with sensitivity and/or specificity equal to 1.0 or 0.0 are not shown here.

Work-flow and economic costs would be huge. Above all, the women would lose confidence with the screening program. The graph representing diagnostic performance of an examination with 95% sensitivity and 80% specificity as a function of disease prevalence is given in Figure 1.2.

Note that we hypothized a disease prevalence equal to 50% for clinical mammography and to 3% for screening mammography to simplify calculations in this example. In the real world, the disease prevalence in screening mammography is about ten times lower (after the first round, only 0.3-0.5% of incident cancers). Thus, the problems we would have in screening

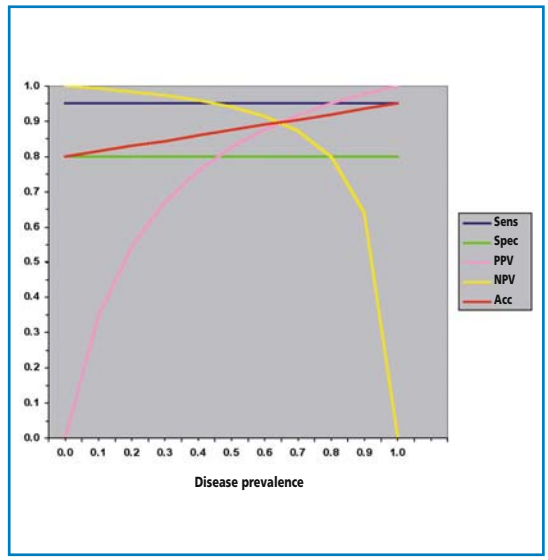


Figure 1.2. Distribution of positive predictive value (PPV), negative predictive value (NPV), and overall accuracy as a function of disease prevalence (constant sensitivity and specificity, equal to 0.95 and 0.80, respectively). Note that with increasing disease prevalence from 0.00 to 1.00 the predictive values change according to two different curves whereas overall accuracy increases linearly from 0.80 (specificity) to 0.95 (sensitivity). At a disease prevalence of about 0.65, overall accuracy, PPV, and NPV tend to be equal (0.89). The likelihood ratios (LRs), not shown here, depend on sensitivity and specificity and are also constant, equivalent to 4.75 (positive LR) and 0.063 (negative LR).

mammography with a relatively low specificity (which could also be accepted in clinical mammography) are also more relevant than those shown by the figures of this example.

Example 1.4. Cardiac CT for diagnosing coronary stenoses. Let us suppose that 64-row multislice CT scanners have a 95% sensitivity and a 95% specificity for clinically significant (≥ 50% reduction in lumen diameter) coronary stenoses. If we were to perform the examination (with intravenous administration of iodinated contrast medium) on 100,000 subjects with a high pretest probability of significant stenoses (80% disease prevalence), we would negate a therapeutic coronary angiography (with stenting of the stenosis) in all the false negative subjects, equal to 5% (4,000 patients). If we were to study 100,000 subjects with a low pretest disease probability (e.g. a screening program for asymptomatic subjects over 65), with the same level of sensitivity and specificity, we would generate 4,750 useless coronary angiographies. This clearly shows that to avoid useless coronary angiographies, even in the presence of high levels of sensitivity and specificity, coronary CT can only be effectively employed with accurate patient selection based on the pretest disease probability defined by means of clinical history and electrocardiogram, stress test, etc. Patients with an intermediate risk (i.e. pretest probability) of significant coronary stenoses (30-70%) are the best candidates for coronary CT. The reader can calculate the predictive values from the data given here. Figure 1.3 shows the graph of diagnostic performance of an examination with 95% sensitivity and 95% specificity as a function of disease prevalence.

A general view of the influence of disease prevalence on predictive values can be obtained looking at the *post-test disease probability* (i.e. the disease proba-

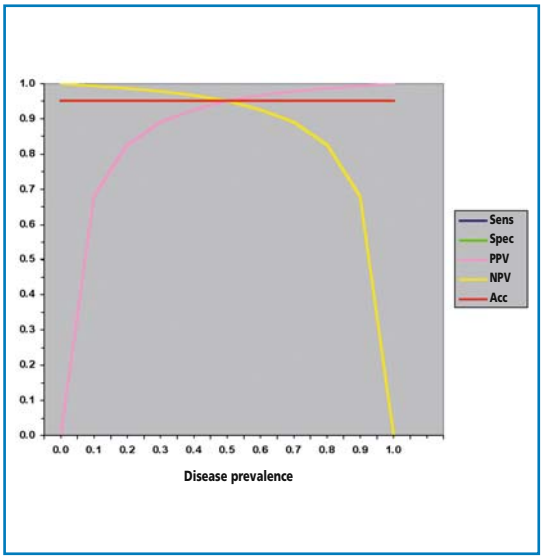


Figure 1.3. Distribution of positive predictive value (PPV), negative predictive value (NPV), and overall accuracy as a function of disease prevalence (constant sensitivity and specificity, both of them being equal to 0.95). Note that: (i) the red line (overall accuracy) overlies the blue line of sensitivity and the green line of specificity; (ii) PPV (pink line) falls drastically when the prevalence falls below 30%; NPV (yellow line) falls drastically when the prevalence goes up above 70%. Likelihood ratios (LRs), which depend on sensitivity and specificity, are also constant, being equal to 19.00 (positive LR) and 0.053 (negative LR).

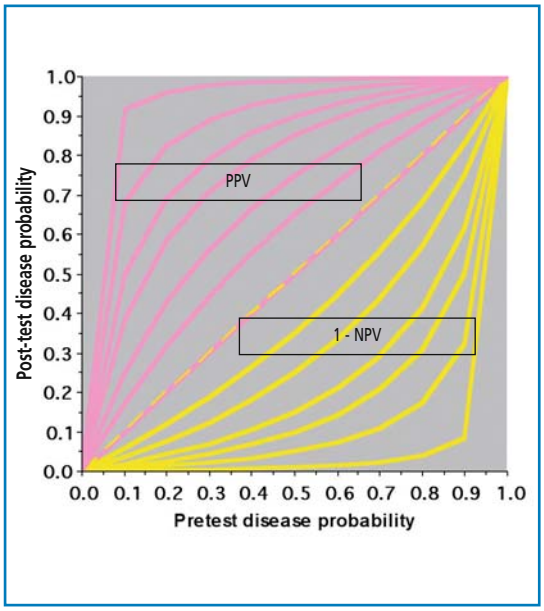


Figure 1.4. Dependence of positive predictive value (PPV) and negative predictive value (NPV) on disease prevalence. The x-axis represents the pretest disease probability (disease prevalence before the examination), the y-axis indicates the post-test disease probability (see text). The curves of PPV (pink lines) and of 1 minus NPV (yellow lines) are given for pairs of values of sensitivity and specificity, both of them being equal to 0.99, 0.95, 0.90, 0.85, 0.75, 0.65, and 0.50 (from the outer to the inner of the graph area). As the pretest disease prevalence increases, both PPV and 1 minus NPV increase (i.e. NPV decreases). The diagonal (mixed pink and yellow) line represents the linear course of the two variables when sensitivity and specificity are both equal to a 0.50.

bility after the examination has been performed) as a function of the *pretest disease probability* (i.e. the disease prevalence in the studied population). Of course, the post-test probability after a positive result is equal to the PPV while the post-test probability after a negative result is equal to 1 minus NPV. A series of curves of PPV and 1 minus NPV are presented in Figure 1.4, each of them obtained for a pair of sensitivity and specificity values.

A first general comment on what we have presented so far is in order. While the predictive values are clearly related to a defined disease ("predictive of … malignant tumor"), sensitivity and specificity may appear to be properties intrinsic to the examination and independent of the disease we would like to confirm or to exclude. This is not the case. Sensitivity and specificity of a radiologic examination do not depend on disease prevalence. However, they must be related to a defined disease. Unfortunately, this relation is frequently omitted or considered implicit. This creates misunderstanding and false expectations for patients and physicians who are non-radiologists. For example, see what we said above in relation to the CT diagnosis of cerebral ischemia and hemorrhage.

Always relate the measures of diagnostic performance to a defined disease

A second comment is also required. Sensitivity and specificity have a different importance according to disease prevalence and severity in the study population. If we study symptomatic subjects (*clinical radiology*), we should try to use examinations with a high sensitivity, even in the presence of a relatively low specificity (this drawback will be compensated for in the following steps of the diagnostic algorithm). In contrast, if we study asymptomatic subjects (*screening radiology*), we should try to use examinations with a high specificity, also accepting a trade-off for sensitivity. In fact, while in clinical radiology the major priority is to diagnose a symptomatic disease (possibly in an advanced stage), in screening radiology the diagnosis of an asymptomatic disease must be balanced by the need of a limited amount of useless diagnostic work-up in the screened population. The consequence is a different way of thinking by the radiologist in the two settings. In clinical radiology, we emphasize even minimal signs as suspicious of disease (especially if related to symptoms), postponing the ultimate diagnosis to the later steps. In screening radiology, we can ignore the minimal signs in order to avoid too high a recall rate.

Two different scenarios: clinical radiology and screening radiology

1.5. Bayes' Theorem, Likelihood Ratios and Graphs of Conditional Probability

The *pretest disease probability* is the probability that a patient has the disease, known before she/he undergoes the examination and the positive or negative result is obtained. In the absence of additional information (personal, family and clinical history, physical examination, and other examinations already performed), the pretest probability is directly equal to *disease prevalence*, i.e. the proportion of the population affected with the disease compared to the entire population. In screening programs, the pretest disease probability is always equal to the disease prevalence in the general population. In clinical radiology, the pretest disease probability is equal to disease prevalence in the general population modified by the selection applied by the referring physician on the basis of medical history and clinical evaluation. In this way we take into account demographic risk factors (age, sex, ethnic group), family history, exposure to other risk factors (e.g. alcohol or smoking), previous and recent medical history, and physical examination.

Bayes' theorem

Bayes' theorem, also called *theorem of subjective probability* or *theorem of conditioned probability,* enables us to calculate – step-by-step in the decisional algorithm – the pretest and post-test probability for a defined disease. It states that the probability that the result of an examination is associated with

the presence or the absence of the disease depends on the pretest probability and on the "power" of the examination. Let us now try to understand what the power of a diagnostic examination is.

The theorem was proposed by the Presbyterian pastor Thomas Bayes (1702-1761) and published posthumously in 1763. Using probabilistic notation, the probability that an event y occurs is defined as P(y); moreover, the symbol "|" means "given that", "if we suppose that", namely that another event conditioning the P(y) has already occurred. Hence, to indicate the probability of the y event, given that the x event has occurred, we write P(y | x). Bayes' theorem states that:

Bayes' theorem

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

where: P(y) is the *a priori probability* of y, P(x | y) is the *likelihood function*; P(x) is the *marginal probability*, that is to say the probability of observing the x event without any previous information and P(y | x) is the *a posteriori probability* of y, given x. P(x | y) / P(x) is the coefficient that modifies P(y) to give P(y | x). It can be shown that P(y | x) is always less than or equal to 1. If the x event is the positive result of a diagnostic examination and we know the pretest disease probability, the theorem allows us to calculate the disease probability (the y event) after having obtained a positive result, i.e. the post-test probability.

Bayesian statistics and frequentist statistics

The concept of *probability* as a degree of our believing that an event happens (*subjective probability*) is the foundation of *Bayesian statistics* and is in opposition with the classic viewpoint of *frequentist statistics*, based on frequencies and proportions (*objective probability*). The Bayesian school has always been a minority among statisticians when compared with the frequentist school. Frequentist methods are today mainly used in medical research, in part due to the possibility of presenting the reliability of an investigated hypothesis as a number (the well-known *p* value). However, especially with regard to the evaluation of diagnostic performance, Bayes' theorem has a basic conceptual relevance, even though sensitivity, specificity, predictive values, etc are managed in the medical literature with classic frequentist statistical methods. The debate between the two schools is still open and animated, in part thanks to the huge calculation power offered to Bayes' supporters by present-day computers.

Odds: a different way of thinking probability

An extended explanation of Bayes' theorem (with the complication given by the possibility of multiple alternative events) is beyond the aims of this book. Here we shall introduce the concept of *odds*. This is probability in a different sense with regard to the usual meaning of frequency as the number of events of interest divided by the whole sample of events. In which sense? Here it is useful to recall *gambling odds*, the probability of winning in a game of chance. In fact the theory of probability was also born in the context of calculations for crap and card games in the 15th and 16th centuries.

Let us consider a practical example. A sample of 10 subjects includes 3 patients affected by a disease. We could say that the frequency of the disease in the whole sample is 3/10, i.e. 0.30, equivalent to 30%. The odds of disease is the ratio between the subjects with the disease and the subjects without the

disease equal to 3/7, i.e. 0.43, or 43%. The odds tell us how many patients with the disease we found for each subject without the disease.

There is a simple mathematic relationship between these two ways of representing the probability (or the risk) of a disease:

if odds = a/b
then frequency in the whole sample = a/(a+b)

Conversely,

if frequency in the whole sample = x
then odds = x/(1-x)

According to Bayes' theorem:

odds of post-test disease = positive LR × odds of pretest disease

This is the equation of a straight line with an angular coefficient equal to the positive LR.

As a consequence, if we have the odds of pretest disease and the positive LR of an examination – which is equal to sensitivity/(1-specificity) – we can calculate the odds of post-test disease. This can be ultimately changed into frequency in the whole sample using the first of the three previous mathematic relations. *In practice, when the positive LR of a test is known, the clinician can change the pretest probability into post-test probability, i.e. into the real diagnostic performance supplied by the test*. Similar reasoning can be proposed for the probability of the absence of disease and the negative LR, which is equal to (1-sensitivity)/specificity.

The logical meaning of likelihood ratios

The logical reasoning behind LRs is now clear. They answer the questions:

- To what extent does the positive result of the test increase disease probability (positive LR)?
- To what extent does the negative result of the test reduce disease probability (negative LR)?

These are two coefficients: when they are equal to 1, they state that the examination does not supply any new information. In fact, post-test probabilities remain equal to the pretest probabilities. Conversely, values of positive LR progressively higher than 1 and values of negative LR progressively lower than 1 indicate increasing levels of diagnostic performance of an examination. In particular, a positive LR higher than 10 implies the examination is ultimately diagnostic for the presence of the disease while a negative LR lower than 0.1 implies that the examination is ultimately diagnostic for the absence of the disease. Intermediate values of LR imply an intermediate degree of diagnostic certainty. *Basically, LRs quantify the power of an examination*.

Likelihood ratios as the "power" of an examination

The reader might suggest that a similar function may also be proposed for sensitivity and specificity. This is partly true, but it is not precisely the same thing. The mathematic mechanism creates a substantial change. We really obtain LRs by a particular mathematic combination of sensitivity and specifici-

ty. However, LRs allow us to change pretest disease probability into post-test disease probability, an important task which uncombined sensitivity and specificity are unable to do.

A simple way to obtain post-test disease probability from pre-test disease probability using LRs is given by the use of a nomogram, a fantastic old mathematic tool used before the advent of computers. It exploits the graphic solution of an equation with multiple variables. *Fagan's Bayesian nomogram* [FAGAN, 1975] changes pretest disease probability into post-test disease probability using a geometric projection, without any need for calculation (Figure 1.5). The slope of the straight line on the nomogram allows us *to graphically see the power of the examination*.

Fagan's Bayesian nomogram

Another way of presenting the relation between pretest and post-test disease probability is the graphs of conditional probability (GPCs) [MALONE AND STAUNTON, 2007]. These graphs supply a visual representation of the change in disease probability obtained using a diagnostic examination in a given clinical setting. The diagnostic performance can be appreciated on the graphs in terms of modification of disease probability for the positive and negative result of the examination at all the points of the range of pretest disease probability and the contribution of different techniques can be evaluated to design efficient diagnostic algorithms. An example is shown for a diagnostic algorithm including D-dimer test, CT pulmonary angiography and indirect CT venography in diagnosing pulmonary embolism [DODD, 2007] (Figure 1.6).

Graphs of conditional probability

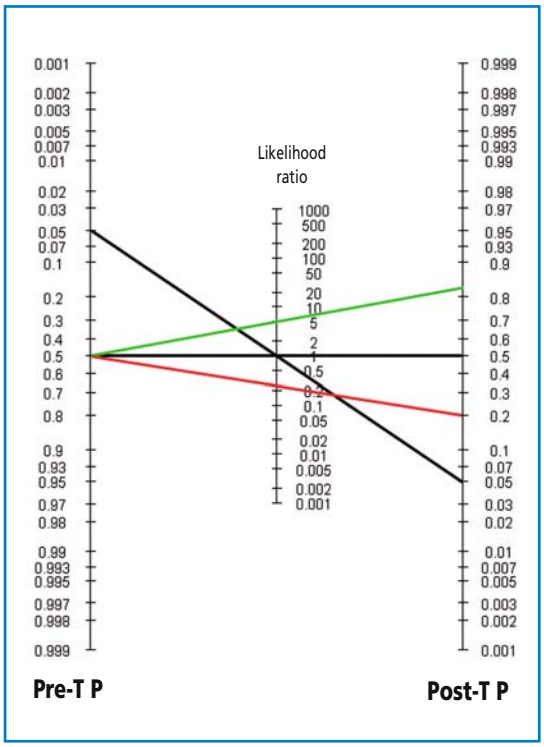


Figure 1.5. Fagan's Bayesian nomogram. The central vertical axis shows the positive and negative likelihood ratio (LR) values, the vertical axis on the left side shows the pretest disease probability (pre-T P), and the vertical axis on the right side shows the post-test disease probability (post-T P). The oblique green line shows how a positive LR equal to +5 changes a pre-test disease probability of 0.5 (i.e. an absolute uncertainty) into a post-test disease probability of about 0.83 (i.e. a relatively high disease probability). The two black lines show how an examination with an LR equal to 1 makes no change to disease probability. The oblique red line show how a negative LR equal to 0.35 changes a pretest disease probability of 0.5 into a post-test disease probability of 0.2. In this way LRs act as angular coefficients of the straight lines designed on the Bayesian nomogram.

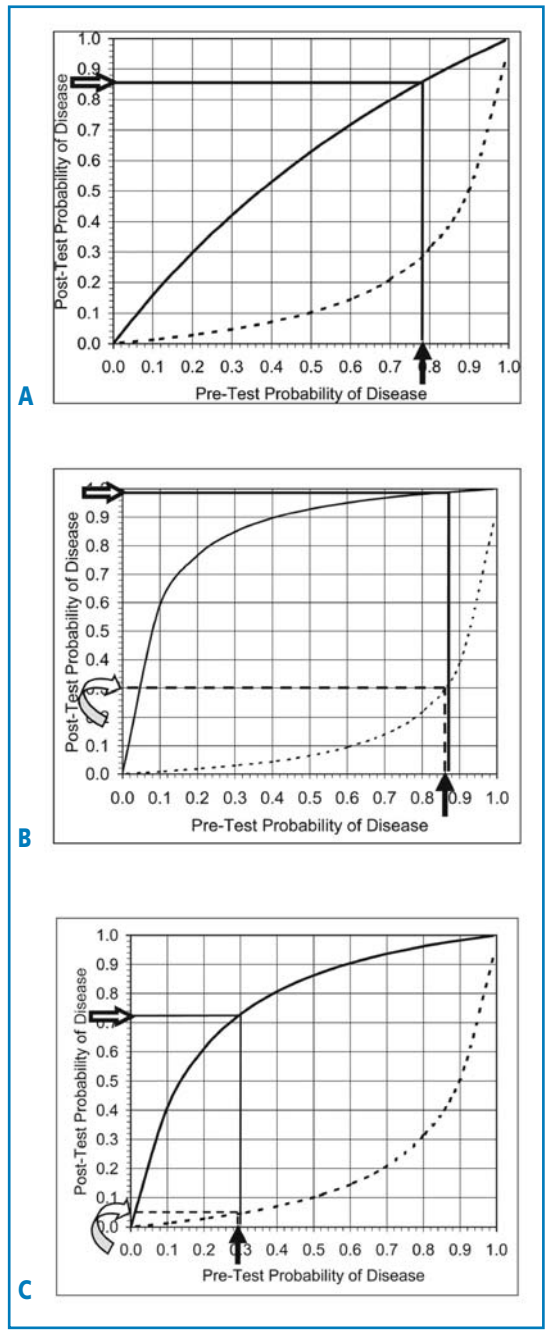


Figure 1.6. Graphs of conditional probability. Diagnostic performance of D-dimer (A), CT pulmonary angiography (B), and indirect CT venography (C) for pulmonary embolism and deep venous thrombosis. Positive result of the examination = solid curve line; negative result of the examination = dashed curve line. For a pretest probability (on the x-axis), the post-test probability of a positive or negative test is derived by drawing a perpendicular line up to the solid line or dashed line, respectively, and then across to the y-axis. For a patient with a high pretest probability of pulmonary embolism, the prevalence is 78% (solid arrow in A). Post-test probability for a positive D-dimer result is 85% (open arrow in A), which warrants further investigation. This post-test probability is then applied as pretest probability to the graph for CT pulmonary angiography (solid arrow in B). If the result is positive, post-test probability is 99% (open arrow in B) and the diagnosis is confirmed. If the result is negative, post-test probability is 30% (curved arrow in B), which does not allow the disease to be ruled out: further investigation is warranted. This post-test probability is finally applied as pretest probability to the graph for indirect CT venography (solid arrow in C). If the result is positive, post-test probability of deep vein thrombosis is greater than 72% (open arrow in C) and diagnosis is confirmed. If the result is negative, post-test probability of deep venous thrombosis is less than 5% (curved arrow in C) and the diagnosis is excluded. From Dodd JD (2007) Evidence-based practice in radiology: steps 3 and 4--appraise and apply diagnostic radiology literature. Radiology 242:342-354 (with permission of the author and of copyright owner [RSNA]).

1.6. Cutoff and ROC Curves

Thresholds and cutoff

In the logical development of our discussion we have left out a relevant aspect. In fact, we supposed that both the radiologist and the pathologist are required to give a dichotomous judgment (yes/no) about the malignancy of the lesion.

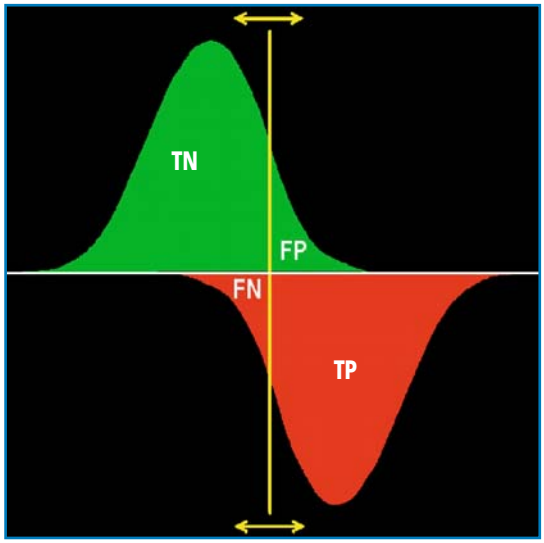


Figure 1.7. Cutoff. Effect of the *cutoff* positioning for a population made up of two equivalent groups of subjects with or without the disease. On the x-axis, a variable (also radiologic) which has higher values in subjects with the disease (inverted Gaussian curve – red area) is compared with the subjects without the disease (Gaussian curve – green area). The red area is inverted only to facilitate the visual evaluation of the cutoff effect. Due to the large overlap of the two curves, the cutoff (yellow vertical line) determines not only the two large fractions of true positives and true negatives, but also the two minor but non-negligible fractions of false positives and false negatives. A cutoff shifted towards the left reduces false negatives but increases false positives, and the opposite occurs with a cutoff shifted towards the right.

However, we know that clinical radiology (and pathology, too) is not made up only of black and white judgments. There is a large *gray scale*, i.e. multiple levels of certainty when we are either more or less in favor of the presence or absence of a disease. This problem is related to the *threshold* we choose for our diagnostic decision, i.e. the *cutoff*. Above the cutoff a radiologic sign is considered predictive of a disease.

The cutoff is an intuitive concept when applied to laboratory blood sample analysis. If the normal upper plasma glucose level is lowered from 120 mg/dL to 100 mg/dL, the subjects with a plasma glucose level from 101 to 120 mg/dL previously considered normal will now be considered abnormal. If a group of these subjects are really abnormal, we would have increased the true positives and reduced the false negatives, with a gain in sensitivity. On the other hand, in the same time the remaining subjects are normal, we would have increased the false positives and reduced the true negatives, thus losing specificity.

Effect of a modified cutoff

If we lower the cutoff, we gain in sensitivity and lose in specificity. If we raise the cutoff, we gain in specificity and lose in sensitivity. This is clearly evident when the variable under investigation is measured on a continuous scale (e.g. blood sample analysis, radiologic lesion sizing in diameter or volume, CT densitometry, bone densitometry, MR signal intensity, evaluation of absolute or percentage contrast enhancement). A graphical representation of the cutoff definition is given in Figure 1.7.

A typical example is the diagnosis of metastatic mediastinal lymph nodes on CT scans on the basis of their size measured as maximal diameter. If we use the classic cutoff which defines nodes larger than 10 mm in diameter as metastatic, we cannot avoid either a fraction of false negatives (metastatic nodes small-

er than or equal to 10 mm in diameter) or a fraction of false positives (non-metastatic nodes larger than 10 mm in diameter). By lowering the cutoff we increase sensitivity but reduce specificity, whereas by increasing the cutoff we increase specificity but reduce sensitivity.

Cutoff optimization

The cutoff could be optimized by choosing the level which minimizes total errors (the sum of false negatives and false positives). However, in clinical practice we adjust – often unconsciously – the cutoff to distinguish normal from abnormal findings in relation to the clinical history and the results of previous examinations which determine the pretest disease probability. For instance, a history of previous malignancy will prompt the adoption of a lower cutoff for the size of a mediastinal node considered suspicious of metastasis. The presence of a deleterious mutation of BRCA1 or BRCA2 genes in women, a relevant family history of breast or ovarian cancer, or the simple personal history of previous breast cancer in the patient prompts the adoption of a lower cutoff reading for the mammography or MR breast examination. In this way a radiologist unconsciously uses Bayes' theorem, increasing sensitivity (and probably losing specificity): s/he believes there is a higher pretest disease probability than would be expected in subjects without these risk factors.

Role of disease spectrum

We can now come back to the matter presented in Section 1.4, i.e. the influence of subject selection on diagnostic performance. We have already stated that, even if the disease prevalence remains unchanged, if the *spectrum* of the subjects with and without the disease changes, both sensitivity and specificity may be significantly altered. A graphical representation of this phenomenon is given in Figure 1.8.

ROC curve

If we do not change the value of variables such as disease prevalence, spectrum of disease severity, etc., can we represent the diagnostic performance of an examination by taking into consideration what happens using different cutoffs? The answer is yes. Note that again, as with the positive LR, we combine "sensitivity" and "1 – specificity", the last term being the false positive rate. As stated above, the positive LR is the ratio between the first and the second term. Now, sensitivity is graphed on the y-axis and 1 – specificity on the x-axis. The points defined by the Cartesian coordinates using different (usually at least five) cutoffs describe the *receiver operator characteristic* (*ROC*) *curve*.

As with ultrasonography and other medical imaging modalities, the ROC curve is the result of a scientific development made in a military context. ROC curves were introduced to optimize the signal detection after the Japanese attack on Pearl Harbor, in order to understand why the radar receiver operators failed to identify Japanese warplanes. Since the 1950s, ROC curves have been used in psychophysiology and have entered the field of statistical methods.

The ROC curve is a tool able to represent the power of a diagnostic examination at virtually all possible cutoffs. In practice, at least five levels are needed to obtain an acceptable curve, as with the BI-RADS® score [American College of Radiology, 2003] (Figure 1.9). The ROC curve intercepts the oblique straight line between the upper left corner and the lower right corner of the Cartesian quadrant. This interception point is the best affordable diagnostic performance with a balance between specificity and sensitivity. However, as stated above, in many situations we might prefer a higher sensitivity with a

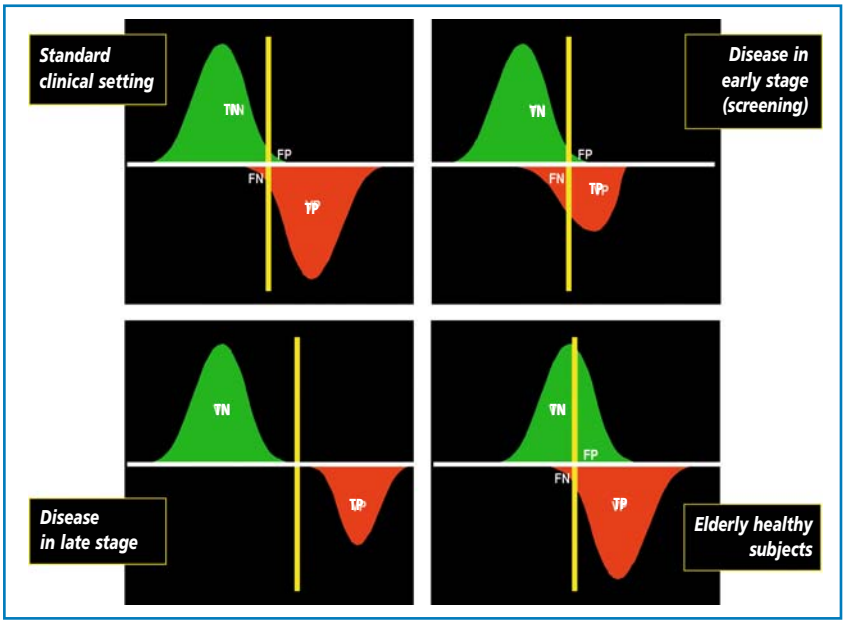


Figure 1.8. Effect of changes in disease spectrum and healthy condition on diagnostic performance. The area under the curve of the distribution of healthy subjects is colored green; the area under the curve of the distribution of the patients affected with the disease is colored red. *Upper left* (standard clinical setting for outpatients): only about 50% of the symptomatic subjects are actually affected by the disease; few subjects produce false negative (high sensitivity) or false positive (high specificity). *Upper right* (screening setting): the patients affected by the disease are lower in number and they also have a disease with a lower mean level of severity. As a consequence the red area under the curve is smaller in size and shifted towards the left with a larger overlap on the green area of the healthy subjects: there are more false negatives resulting in a lower sensitivity and negative predictive value. *Lower left* (clinical setting for in-patients): the mean level of disease severity is higher; the red area is smaller (some patients have died) and shifted towards the right; by shifting the cutoff towards the right, we can distinguish perfectly between patients with disease and healthy subjects (no false negatives or false positives). *Lower right* (changed spectrum of healthy subjects): a more aged healthy population shifts the green area to the right, producing more false positives and a lower specificity and positive predictive value.

tradeoff in specificity or vice versa. Figure 1.10 shows a series of ROC curves with increasing diagnostic performance.

ROC analysis in radiology

A relevant application of the ROC curve in radiology, in the setting of the diagnosis of a defined disease, is the comparison between different imaging modalities or different approaches (e.g. new or old techniques), or different readers (e.g. with extensive or limited experience) for a single imaging modality. This comparison can be performed on the same sample of patients or in different samples of patients. It is noteworthy that in the latter case (different samples of patients), the comparison between the two AUCs gives a result equivalent to the application of the Mann-Whitney *U* test, the typical non-parametric test for unpaired data (see Chapter 5). This shows that apparently different aspects of biostatistics are actually connected by a logical-mathematic relation.

Understanding one part of the image helps us to understand another part of the image. In the end the whole picture will seem less complicated in comparison with our first impression due to the initial difficulties.

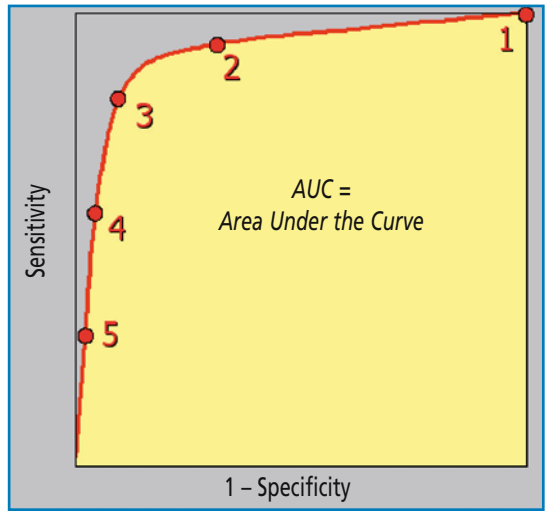


Figure 1.9. ROC (receiver operator characteristic) curve. The graph representing the relation between sensitivity and 1 – specificity (i.e. the false positive rate) enables the quantification of the power of a diagnostic examination as the *area under the curve* (AUC). In the example, five levels of cutoff are indicated on the ROC (like the BI-RADS® score system). The Cartesian coordinates of point 5 are the sensitivity and 1 – specificity we obtain considering only the findings scored as BI-RADS® 5 as positive with low sensitivity and very high specificity, and so on until point 1 with 100% sensitivity and 0% specificity (BI-RADS® 1 is the score for a completely normal examination).

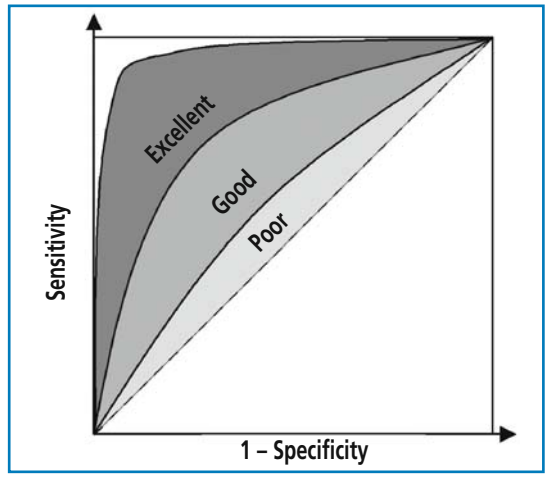


Figure 1.10. ROC (receiver operator characteristic) curves. Examinations with larger areas under the ROC curve have diagnostic performances higher than those with smaller areas under the ROC curve. Only the ROC curves above the oblique straight line between the upper right corner and the lower left corner supply useful diagnostic information. The different degree of gray indicates areas of poor, good, and excellent ROC curves.

References

American College of Radiology (2003) ACR breast imaging reporting and data system (BI-RADS): breast imaging atlas. Reston

Dodd JD (2007) Evidence-based practice in radiology: steps 3 and 4 – appraise and apply diagnostic radiology literature. Radiology 242:342-354

Fagan TJ (1975) Nomogram for Bayes theorem. N Engl J Med 293:257

Malone DE, Staunton M (2007) Evidence-based practice in radiology: step 5 (evaluate) – caveats and common questions. Radiology 243:319-328

Sardanelli F, Giuseppetti GM, Panizza P et al (2004) Sensitivity of MRI versus mammography for detecting foci of multifocal, multicentric breast cancer in fatty and dense breasts using the whole-breast pathology examination as a gold standard. AJR Am J Roentgenol 183:1149-1157

Sobue T, Moriyama N, Kaneko M et al (2002) Screening for lung cancer with low-dose helical computed tomography: anti-lung cancer association project. J Clin Oncol 20:911-920

2

Variables and Measurement Scales, Normal Distribution, and Confidence Intervals

> Science is built up of facts, as a house is with stones.
> But a collection of facts is no more a science
> than a heap of stones is a house.
>
> JULES HENRI POINCARÉ

The dilemma between sensitivity and specificity noted by the choice of threshold arises from the intrinsic variability of biologic phenomena, both at the cellular level and the organ level in the human body in the presence and absence of pathologic processes. When one measures the same hallmark in a sample of individuals there always appears a *spectrum of values* which is a more-or-less wide numerical set characterizing that sample for the measured hallmark. It is not by chance that in Figure 1.6 we used a bell-shaped curve to represent the set of possible values of the measured variable. Such curves indicate that the variable may take *all* the values within them, and that the most frequently observed values correspond to the central part of the curves.

In other circumstances the variable of interest may assume only *qualitative* values. This happens, for example, in the presence or the absence of a radiologic sign: if we study a sample of *n* individuals, only a part of them will show that sign.

Variables and distributions

The object we are measuring is termed a *variable*. The values that it may take depend on a mathematical law called *distribution*. One of the goals of statistics is the characterization and representation of variables and their distributions. In this chapter we will discuss the main types of variables and the essential elements of *Descriptive Statistics* (statistics which *describes* the characteristics of the data). Therefore, we will see the main features of *Gaussian distribution*.

The reader should note that the subjective perception of a radiologic sign also has its own variability: it can be different for two or more observers in the same study and different for the same observer in different conditions. This concerns a special topic, the *reproducibility* of a diagnostic study, which will be given particular attention in Chapter 7.

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

2.1. Variables and Measurement Scales

We define variable a feature that can be observed and/or measured and that may take at least two different values. Common synonyms (also in this book) include *characteristic* and *quantity*. *A variable is a kind of container that can contain any type of information, but the representation and processing of this information depend on the type of the data.*

Statistical analysis mainly depends on the variable type

An important point is the subtle difference between the variable, its type and the measurement scale used to represent it. The *measurement scale* is dependent on the values the variable may take and the procedure (instrumental measurement or subjective judgment) with which these values are obtained. Changing the measurement scale may switch the variable from one type to another. For example, let us consider the degree of stenosis of the carotid arteries. We can indicate this variable with a numerical value that represents the percentage of occlusion; otherwise, we can visually distinguish the degree of stenosis as mild, moderate or severe. In both cases, the variable of interest is the degree of stenosis, but in the former case we have a measurement scale ranging from 0% to 100%, while in the latter case we may use only three categories. As we shall see, this change in the measurement scale makes the variable (the degree of stenosis) switch from the continuous type to the ordinal type. *The measurement scale therefore clearly defines the type of variable.* For this reason, some authors believe that the classification we propose in Sections 2.1, 2.2 and 2.3 can be attributed to the measurement scale, with no distinction between the type of variable and the measurement scale. Although in practice the two concepts are equivalent, in some circumstances the difference between the type of variable and the measurement scale is evident.

Link between variable type and measurement scale

In what follows, we report a brief summary of the different types of variables and measurement scales [Siegel and Castellan, 1992]. The difference between types of variables is very often quite subtle and at a first glance this difference may not be so clear. We therefore invite the reader to pay close attention, since the statistical analysis strongly depends on the type of the variables of interest. The first important distinction is between *categorical* variables and *numerical* variables.

2.1.1. Categorical Variables

Nominal variables

Categorical variables are variables whose values define categories, i.e. characteristics of the individual that have no natural order. Typical examples are race, gender, imaging technique, radiologic subspecialty, etc. For these examples, the values they may take represent only *names* (Asian, female, MR imaging, interventional radiology, etc.) and for this reason these variables are also called *nominal*. A special case of categorical data is the dichotomous variable, like the result of a diagnostic study in "positive" or "negative".

Ordinal variables

In some cases, for example in the BI-RADS® score system for reporting mammograms [American College of Radiology, 2003], there is an intrinsic order in the data, even though the difference between different scores cannot be quantified. In these cases the variable is called "ordinal". Another example is TNM cancer staging [UICC, 2002].

Ranks

The statistical analysis of ordinal data is often performed by converting each category into *ranks*, i.e. with the association of progressive numerical values which are easier to manage. Typically, a sequence of integer numbers (1, 2, 3…) is assigned to the various values of the variable. The radiologist's judgment may, for example, be expressed using the BI-RADS® measurement scale (from 1 to 5) instead of negative, benign, probably benign, suspicious abnormality, highly suggestive of malignancy.

Converting the ordinal variables into ranks is the conceptual link between the categorical and numerical variables. The statistical analysis for the latter is generally more powerful.

2.1.2. Discrete Numerical Variables

Discrete numerical variables may take only a limited number of numerical values. Generally, they regard countable values such as age, the number of lesions, etc.

Difference between discrete numerical and ordinal variables

The difference between discrete numerical variables and ordinal variables is an important one. For example, let us focus our attention on number of malignant lesions (discrete numerical variable) and tumor staging (ordinal variable): four malignant lesions are twice as many as two malignant lesions, but tumor stage II cannot be considered as twice the value of tumor stage I.

Interval variables

With discrete numerical variables the difference between two consecutive values is constant (e.g. Hounsfield units in CT)[1] and this difference represents an interval. For this reason, these variables are also known as *interval* variables.

2.1.3. Continuous Numerical Variables

Dimensional measurements are continuous variables

Continuous numerical variables may take an infinite number of values which are generally obtained by direct or indirect instrumental measurement. Due to the possibility of being expressed with an arbitrary number of decimals, these variables may, in theory, take every value in a given interval. In radiology, typical examples are lesion size, MR signal intensity, organ volume, artery diameter, etc. These variables are often measured by dedicated computational tools within the processing units.

In some circumstances, discrete variables can be managed as if they were continuous variables, provided the sample has many different values. For example, let us consider the age of a sample of 30 individuals, expressed in years: if the age distribution covers a range from 20 to 80 years, then this variable can be considered as continuous, even if it is a discrete variable. To do the same with children, age needs to be expressed in months instead of years, and with a sample of newborns, in days instead of months. Therefore, choosing the right measurement scale is important for the statistical analysis one wants to perform. Moreover, the opposite procedure to the one just explained can also be performed. Indeed, a continuous variable can be con-

Continuous variables may also be considered discrete

[1] With Hounsfield units, the difference between the electron density of tissues and that of water is divided by the water electron density and then multiplied by 1000. This approach distributes image contrast over a wide range.

sidered discrete if we divide its value interval into two or more subintervals. These subintervals may have the same amplitude (the continuous variable becomes an interval variable) or different amplitudes (the continuous variable becomes an ordinal variable). For example, the NASCET criteria [NASCET, 1991] for the classification of carotid artery stenosis uses the following categories:

≤ 29% = mild stenosis;
30%-69% = moderate stenosis;
≥ 70% = severe stenosis.

In this case, a continuous variable like percentage occlusion is converted into ordinal data thanks to the switch from one measurement scale to another.

2.1.4. Measurement Scales

Analogies between variable types and measurement scales

The reader should have noted that the different data types have been defined in relation to the possible values they may take, i.e. on the corresponding measurement scales. As stated above, these are not independent concepts. The classification of measurement scales can be done in the same way as the classification for variables, as Table 2.1 shows.

In medicine all types of variables and measurement scales are constantly in use. A relevant part of the radiologist's interpretation consists of converting

Table 2.1. Measurement scales

Type	Definition	Hallmarks	Examples
Qualitative	Nominal or categorical	Absence of a hierarchy or order within categories	Positive/negative (dichotomous variable); race, gender, imaging technique, radiologic subspecialty
	Ordinal or ranked	Presence of a hierarchy within categories but the difference between two consecutive values cannot be quantified	BI-RADS® score for reporting breast examinations
Quantitative	Interval	Constant interval between two consecutive values without a starting zero point; the variable may take positive and negative values; it does not allow proportional calculation	Electron density in computed tomography (Hounsfield units), temperature measured in degrees Celsius, T-score in bone densitometry
	Rational	Constant interval between two consecutive values with a starting zero point; the variable may take only positive or negative values; it allows proportional calculation	Heart rate, signal-to-noise ratio

continuous data into a categorical evaluation, up to defining the examination as positive or negative for the presence of a given disease.

As stated above, recognizing the data type being analyzed is highly relevant because, while under certain conditions numerical variables may be manipulated with parametric statistical techniques (see Chapter 4), categorical data must always be analyzed with non-parametric methods (see Chapter 5).

2.2. Gaussian Distribution

In the previous section we learned about the classification of variables. Now we shall introduce an extension of what was stated above regarding continuous variables. The concept of a *distribution* is very intuitive. A complete explanation of all possible distributions (both continuous and discrete) is beyond the aims of this book. To this end the interested reader may consult specialized texts [Soliani, 2007]. However, the reader should pay particular attention to this section given its importance for parametric statistics.

How Gaussian distribution is built

Let us suppose that a sample of 50 males, aged 20-50 years and without cardiovascular diseases, undergo abdominal CT; for each subject the diameter of the suprarenal abdominal aorta is measured. Table 2.2 shows the results.

In this sample the values of the abdominal aortic diameter are very close to 30 mm, with a minimum of 26.5 mm and a maximum of 33.4 mm. Data are expressed using only a decimal place with the 50 individuals being distributed over a range of 33.4-26.5 = 6.9 mm.

The observation of the raw data cannot provide a complete evaluation of all the information contained. A more suitable way of handling the data is to divide the observed value interval into subintervals and to determine how many measurements lie in each. For example, we may consider the number of aortas

Table 2.2. Aortic diameter of a sample of 50 healthy individuals

No.	Diameter (mm)	No.	Diameter (mm)	No.	Diameter (mm)
1	29.8	19	30.3	37	32.5
2	30.2	20	31.0	38	33.4
3	30.1	21	30.5	39	26.5
4	31.2	22	29.6	40	27.4
5	28.6	23	32.3	41	30.4
6	29.7	24	27.9	42	30.5
7	30.5	25	28.5	43	31.0
8	30.9	26	28.9	44	29.6
9	31.2	27	31.4	45	29.8
10	29.4	28	31.6	46	33.1
11	29.2	29	30.1	47	30.0
12	29.9	30	30.6	48	30.1
13	27.5	31	30.7	49	29.8
14	27.2	32	29.7	50	30.1
15	31.8	33	29.9		
16	32.2	34	29.3		
17	30.2	35	30.1		
18	29.9	36	30.2		

Table 2.3. Number of aortic diameters in each subinterval

Interval (mm)	Counts
26.0-26.9	1
27.0-27.9	4
28.0-28.9	3
29.0-29.9	13
30.0-30.9	17
31.0-31.9	7
32.0-32.9	3
33.0-33.9	2
Total	50

whose diameter lies in the subintervals 26.0-26.9 mm, 27.0-27.9 mm, 28.0-28.9 mm, and so on. Table 2.3 summarizes the number of *counts* within each subinterval.

There are no aortic diameters outside the interval 26.5-33.4 mm, while a substantial proportion (30/50, 60%) of them lie within the two central subintervals. The next step is to report the data of Table 2.3 as a graph, as shown in Figure 2.1.

Histogram

This type of graph is called a *histogram* and it provides an immediate insight into the information contained in Table 2.2. Indeed, the more populated subintervals are the central ones and the number of diameters diminishes rapidly when moving away from the centre. The subdivision into subintervals is arbitrary and depends on the sample size; however, a good compromise between number of subintervals and counts is advisable.

Let us now suppose we increase the sample size from 50 to 200. The reader may easily see that in this new case many more subintervals with a reduced amplitude may be taken into consideration. When instead of considering only

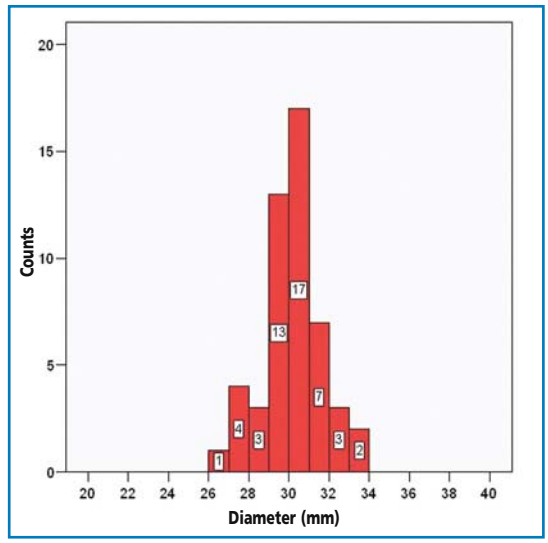


Figure 2.1. Histogram of the abdominal aortic diameter. The x-axis reports the subdivision into subintervals proposed in Table 2.3. The y-axis indicates the number of aortic diameters within each subinterval. The reader should note that the x-axis does not start at zero.

a sample we consider the whole *population*[2] of males aged 20-50 years without cardiovascular diseases in a given geographic area, we would reduce the subinterval amplitude to such a point that the histogram will appear as a continuous[3] bell-shaped curve, as showed in Figure 2.2.

Difference between histogram and distribution

The curve in Figure 2.2 is called *population distribution* and represents *a limit case never encountered in practice*. One of the most interesting aspects of statistics is its capacity to extrapolate information obtained from a sample (necessarily limited) to the entire population. This aspect is the goal of *inferential statistics* and will be explained at length in the following chapters.

From the sample to the population

When analyzing data from more-or-less limited samples, the term of reference will always be histograms. Often the word "distribution" is also used for limited samples, but it is important to stress the terminological difference: *histograms for samples; distributions for populations*.

Histograms for samples, distributions for populations

The reader may be wondering why patients with cardiovascular diseases and those younger than 20 years and older than 50 years of age were excluded. This was done so that the trend in aortic diameter should be a random variable barely dependent on other factors (age, gender, diseases). Later we will come back to this feature.

Random variable

The distribution of a random variable is *always* bell-shaped, as shown in Figure 2.2. From a rigorous point of view, this curve is described by a mathe-

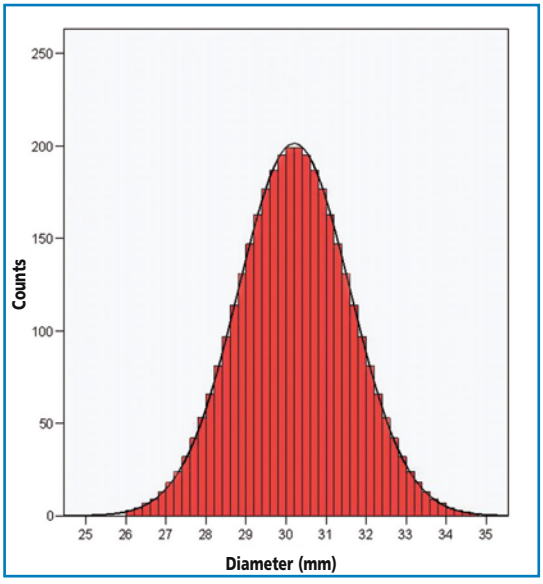


Figure 2.2. Histogram of abdominal suprarenal aorta diameter of the entire population. Note the bell-shaped curve which represents the limit condition when the amplitude of each subinterval becomes zero.

[2] In statistics, the population is an ideal set made up of an infinite number of units. However, in *medical statistics* population stands for a real set of individuals (persons) who have a common characteristic such as a defined nationality, the whole set of patients with myocardial infarct or with prostate cancer, or breast cancer, or patients studied with a certain contrast agent, etc.

[3] From a mathematical point of view, the histogram becomes a real continuous curve only when the interval amplitude is reduced to zero.

Gaussian distribution is also defined *normal*

matical function introduced by Karl F. Gauss (1777-1855) who started his investigation from the geodetic measurements of the German State of Hanover. This function was then used by Gauss to describe the motion of heavenly bodies. Francis Galton (1822-1911) then proposed its use to describe many natural phenomena, arguing that this distribution was the "*norm*" in nature. For this reason Gaussian distribution is also defined "*normal*"[4].

No formal demonstration of the fact that a random variable always has a normal distribution is available. Indeed, this is a *principle*, i.e. a law always empirically verified and never contradicted. The reverse of this law is also used to verify the randomness of a given variable: in practice, if we have a statistical sample from which we measure a continuous variable, drawing up the corresponding histogram and verifying that it has an almost Gaussian shape is enough to conclude that the variable is random[5].

Mean and standard deviation of a probability distribution

The population distribution (built with counts) may be converted into a *probability distribution*, expressed by a mathematical function which allows us to calculate the probability that the measured variable lies within a given interval. This function is[6]:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.1}$$

where μ indicates the curve centre, i.e. the x-axis point where the distribution takes its maximum value; σ is a width parameter, such that if σ is small, the curve is narrow and high, whereas if σ is large, the curve is short and wide[7].

The main feature of Gaussian distribution

We have decided to introduce the mathematical equation of the normal curve in order to discuss one of its most important features. With this function, 95% of the observations lie within the interval [μ – 1.96σ, μ + 1.96σ] (Figure 2.3). In practice, if we measure a characteristic (variable) of the whole population, 95% of them would have a value lying within this interval. Therefore, the probability a given individual has a measured value within [μ – 1.96σ, μ + 1.96σ] is just 95%. Only the remaining 5% of individuals will have a value of x within the two tails of the curve. As we shall see in Section 2.6, the definition of the confidence intervals is based on this feature.

Asymmetric distributions

Now let us reconsider the previous example of the abdominal aorta. If we add children to the sample, we introduce a series of values lower than that reported in Table 2.2 and the left tail of the histogram becomes closer to zero. If, instead, we insert adult females, another maximum in the histogram will be

[4] The terms *Gaussian* and *normal* are synonymous.

[5] Note that the frequent occurrence of normal distributions in biologic phenomena is due to their genesis being made up of a large number of factors (of which we know only a small part). Such factors tend to both increase and decrease the value of the variable, thus determining the substantial randomness of the result.

[6] This function gives the probability p(x) that the measured variable lies within the interval [x, x+dx].

[7] 2σ is the distance between the two curve points where the concavity changes its sign. It represents the curve width in a point placed at 60.7% of the maximum (see Figure 2.3).

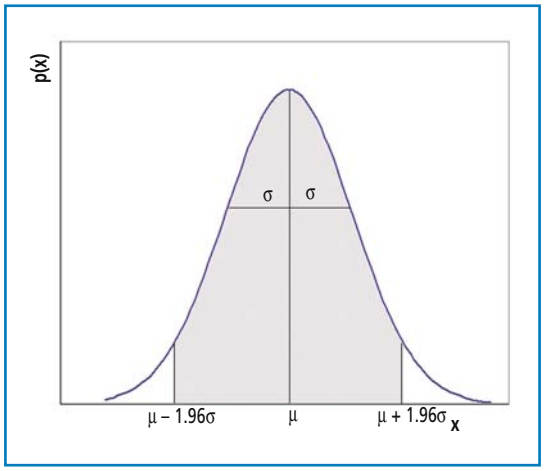


Figure 2.3. Gaussian probability distribution centered at μ and with width 2σ. The probability an individual of the population has x within the interval [μ – 1.96σ, μ + 1.96σ] is 95%.

produced at a lower value than that observed in the adult males at about 30 mm. Similarly, if we insert patients with cardiovascular diseases into the sample, we have a percentage of individuals with a higher aorta diameter who tend to push the right histogram tail to higher values. In these three cases, the histogram will appear *asymmetric* and, as stated above, this means that the measured variable is no longer random.

The Gaussian probability distribution is symmetric about μ and its width depends on σ (see Fig. 2.3). Without going into the mathematical demonstration, it is easy to see that μ *coincides with the mean and* σ *coincides with the standard deviation of the variable we are measuring in the population.*

To further elucidate this feature, let us consider once again the example of the abdominal aorta and proceed step by step. In the data presented in Table 2.2 the aortic diameter tends to lie at about 30 mm and the arithmetic mean (which will be defined in the next section) confirms this trend, being equal to 30.1 mm. However, the sample of Table 2.2 includes only 50 individuals instead of the entire population. The only way to obtain the *real* mean of the abdominal aorta diameter is to measure this variable in the entire population; but this is practically impossible. However, as stated above, the probability distribution is ideally built only for the entire population. It is therefore clear that the maximum point of the histogram of samples with progressively increasing size, which coincides with μ, slowly becomes the mean of the entire population. Similarly, the standard deviation (which will be defined in the next section) of the data in Table 2.2 is a measure of the histogram width and it will become the standard deviation of all the population as the sample size increases. Since the trend of the histogram is to appear as a normal curve with width σ, the standard deviation will clearly become equal to σ.

Sample mean and population mean

The normal probability distribution is completely defined by the two parameters μ and σ: once we know their values, the curve is obtained with Equation 2.1 and it is unequivocally defined. Two distributions with different μ values are displaced from each other on the x-axis, whereas if they have different σ values the two curves have different amplitudes and widths. In the example of the abdominal aorta, the diameter distribution in the adult female population is

Gaussian distribution only depends on mean and standard deviation

probably centered about a lower value, with a partial overlapping with the curve representing adult males.

Standard normal distribution

Figure 2.3 clearly shows that each point on the x-axis, i.e. each value of the measured variable, may be expressed in terms of the distance from the mean $(x - \mu)$. For example, the point $x = \mu + 1.96\sigma$ is placed at a distance equal to $x - \mu = 1.96\sigma$. Since this statement holds for each pair of parameters (μ, σ), they can be made independent by considering the variable:

$$z = \frac{x - \mu}{\sigma}$$

It can be demonstrated that if x is a random variable and, therefore, has normal distribution, then z is also a random variable with normal distribution (called *standard normal distribution*), but unlike x it has a *mean always equal to 0* and a *standard deviation always equal to 1*. The z distribution graph is shown in Figure 2.4.

In practice, *for each random variable* it is always possible to build the corresponding standard normal distribution which is always the same regardless of the starting variable (aorta diameter, size of a lesion, renal volume, etc.). Since x, μ and σ have the same unit of measurement (mm in the case of the aorta diameter), then z is a pure number, i.e. it has no unit of measurement. For all these reasons, *the standard normal distribution is universally used in all inferential statistics*. The reader should note that, with respect to Figure 2.3, in Figure 2.4 the x on the x-axis has been replaced with z, and the p(x) on the y-axis has been replaced with p(z). The two points $x = \mu \pm 1.96\sigma$ become $z = \pm 1.96$ and the interval [-1.96, 1.96] contains 95% of the observations. In terms of probability, the z value for each individual in the population has 95% probability of lying within this interval.

It should be stressed that a deep mathematical understanding of this section is not fundamental. For all practical purposes a number of easy-to-handle data tables are available.

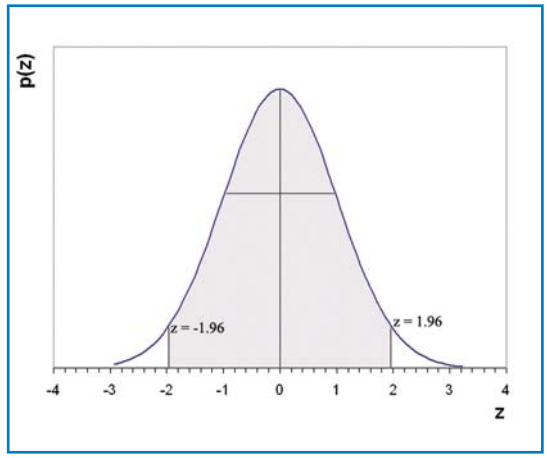


Figure 2.4. Standard normal distribution.

2.3. Basics of Descriptive Statistics

As stated in the Introduction, the goal of *descriptive statistics* is to describe the data of a sample. The word "sample" identifies a set of statistical units (often, in medicine, humans, but sometimes organs, anatomic structures or lesions) *extracted* from a population with one or more features. For example, the population may be all the members of a country (epidemiology), the entire set of newborns (neonatology), all cancer patients (oncology), all patients with a clinical indication for a certain radiologic examination (radiology) etc. Although in all these cases the size of the population is not actually infinite, this size is so large that we can treat it as if it really were infinite.

The population

The population from which the sample is extracted has a precise distribution (not necessarily a normal distribution) and its features reflect that of the sample. If, for example, in a sample of *n* pulmonary nodules at CT screening we observe a high fraction of benign lesions, this suggests that in the entire population of nodules the observed variable (the fraction of benign lesions) will have a value[8] close to that obtained in the sample, and the larger the sample is, the more valid such a statement is.

The population always has an unknown true value

The example just reported introduces a fundamental concept: *random sampling*. This consists of extracting the sample from the population in a totally random way, without selection or influence of any type[9]. Otherwise, the sample characteristics do not reflect those of the population: in this case, the study is affected by systematic distortion or *bias* (bias will be discussed in Chapter 9).

Random sampling

Descriptive statistics is extremely broad and a complete discussion goes beyond the aims of this book. Instead, we shall introduce the most important and most used parameters.

2.3.1. Measures of Central Tendency

Measures of central tendency are parameters that provide information about the position of the distribution.

The first is the arithmetic mean, often simply called *mean*. Let x indicate a continuous variable and let $\{x_1, x_2, \ldots, x_n\}$ be a sample of x. The arithmetic mean m is defined as:

Arithmetic mean

$$m = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (2.2)$$

and is the ratio between the sum of all the measurements and the size of the sample. The reader will have noted that we used the Latin letter "m", in contrast to the previous section where we used the Greek letter "μ".

"m" and "μ"

[8] The mean value of the population is often called *the true value*, because it exists even if we do not know what it is.

[9] It is not by chance we use the verb "to extract", derived from the extraction procedure of a balloon from a box.

This notational difference is generally used to distinguish the estimation of the mean (calculated from the sample) from its true value (that of the entire population).

The arithmetic mean takes into account all the sample values and it is strongly dependent on possible extreme, isolated (*outlier*) data typical of asymmetric histograms. At this point the reader should recall that the mean of ordinal variables cannot be calculated. If, for example, we use the BI-RADS® classification (0, 1, 2, 3, 4, 5 and 6) to describe a sample of mammalian lesions, we could be tempted to calculate the mean sample value. However, although it is possible from a mathematical point of view, we would obtain an absolutely nonsense value. A mean score equal, for example, to 3.4 is not correctly interpretable, because we are unable to quantify the difference between two consecutive scores.

The mean of ordinal variables cannot be calculated

Another commonly used measure of central tendency in statistics is the *median*. It is not *calculated* from the sample data, as with the mean, but it is defined as the value that divides the sample in two sub-samples with the same size, as defined by the following operating procedure:

Median

1. All values should be reorganized in ascending or descending order;
2. For an odd value of n, the median coincides with the central value;
3. For an even value of n, the median is the arithmetic mean of the two central values.

Let us consider a practical example. Let the following samples be the age (expressed in years) of two different groups made up of 15 patients:

18, 18, 23, 27, 32, 35, 36, 38, 38, 42, 47, 51, 52, 56, 57 Group I

18, 18, 23, 27, 32, 35, 36, 38, 38, 42, 47, 51, 52, 86, 87 Group II

The two groups are almost identical samples apart from the last two values that are markedly different (outliers). We obtain:

Outliers

mean = 38, median = 38 (years) Group I

mean = 42, median = 38 (years) Group II

Because of an odd value of n for both groups, the median coincides with the central value, in such a way that seven values are lower and seven values are higher than the median. The mean is 38 years for Group I and 42 years for Group II: it is clear how the mean is influenced by the two extreme values of Group II (86 and 87 years), unlike the median which in contrast is the same value for both groups. This effect depends on the definition of the mean, which is calculated using all the sample data, while the median is a *position* index, placed at the middle of an ordered series of values.

Mode

Lastly, let us introduce another measure of central tendency: *mode*. The mode is the most frequent value of the sample, i.e. the value observed with the highest frequency. It is not necessarily a single value (in the last example the values 18 and 38 are observed twice). In the event of more than one mode the

sample is called *multimodal*[10]. Generally the mode is rarely used, in part because it may be very far from the distribution centre. However, it has great conceptual significance: when we are dealing with a nominal scale of measurement, the mode is the only index of what actually happens in the sample. We usually think of it as the most frequently observed category. In other words, when we have nominal data, asking which of two or more categories is more frequent is the same as defining the mode.

In order to clarify the relationship between mean, median and mode let us consider the following example.

Example 2.1. Let us consider the size of the mediastinal lymph nodes studied by CT in a sample of patients with lung cancer. The sample may contain many small lymph nodes (healthy, inflamed and metastatic) and a progressively decreasing number of enlarged lymph nodes (these are especially, but not only, metastatic lymph nodes). An example of the possible population distribution from which the sample could be extracted is shown in Figure 2.5A. For the purpose of comparison, the distribution of lymph node diameter in the healthy population is also shown (Fig. 2.5B). Note that the relative position among the three indices depends on the symmetry/asymmetry of the distribution: they coincide only in the event of symmetrical distribution.

Example 2.1 shows the importance of calculating both the mean and the median of a sample. By comparing them we derive a fundamental bond for

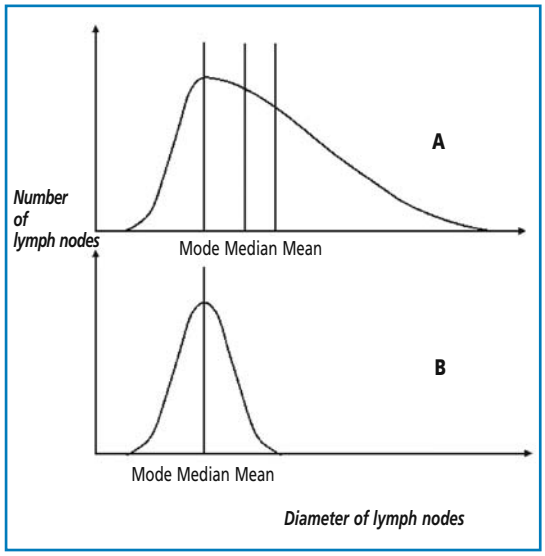


Figure 2.5. Mean, median and mode. The x-axis depicts the diameter of lymph nodes and the y-axis indicates the number of lymph nodes in the population with lung cancer (**A**) and in the healthy population (**B**). The reader should note the difference between mean, median and mode for asymmetric distribution (**A**) whereas the three indices coincide for the symmetric curve (**B**).

[10] This can be adequately demonstrated with the abdominal aorta example introduced in Section 2.2. If the population from which the sample is extracted includes both males and females, the distribution has two maximum values: a maximum corresponding to the mean diameter of the abdominal aorta in males, and another corresponding to that in females.

When the arithmetic mean is not suitable as a measure of central tendency

applying the methods of parametric statistics: the symmetry or asymmetry of the population distribution[11]. *If the difference between mean and median is too large, the median should be used as the measure of central tendency.*

2.3.2 Data Spread about the Measurement of Central Tendency: Variance and Standard Deviation

In the previous section we introduced some measurements of central tendency which, when calculated on a sample, provide information about the *position* of the distribution. If we measure the same variable in two samples extracted from different populations (with different distributions), the two means will tell us how much the two centroids[12] are separated. However, we do not have any information about the *shape* of the distribution, i.e. the way data *spread* about the distribution centroid. The reader will undoubtedly have realized that what we are searching for is an index that measures the distribution width.

Distribution shape indices

Let us reconsider the example of the abdominal aorta that we introduced in Section 2.2. The arithmetic mean diameter is 30.1 mm, the minimum value is 26.5 mm and the maximum value is 33.4 mm. Both the minimum and the maximum values define the *range* of the observed values, but do not provide information on *what happens within* this interval: data could be distributed in several ways, but we know that most of the values are grouped about the mean (see Fig. 2.2).

The starting point is to calculate the distance between each sample value and the mean. Let x_i be the i-th sample value and $\bar{x}$ be the sample mean. The distance between them is the difference $d_i = x_i - \bar{x}$. The difference d_i is also called "*residual*" and it is a positive value when $x_i > \bar{x}$, a negative value when $x_i < \bar{x}$ and it is zero when $x_i = \bar{x}$. For a well known theorem, the sum of all the residuals is zero[13], so we need a different indicator. One possibility is to use the square of the residuals, $d^2_i = (x_i - \bar{x})^2$. The *variance* is defined as:

Variance

$$s^2 = \frac{\sum_{i=1}^{n} d_i^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

and is calculated as the sum of the square of all residuals divided by the *degree of freedom*[14], i.e. (n-1). For variance (s^2) we also used a Latin letter, in order to differentiate the variance calculated in the sample[15] and the variance of the entire population, expressed with σ^2.

---

[11] Actually, verifying the symmetry of the distribution is not enough. If we want to use the methods of parametric statistics we need to check for normal distribution.
[12] We cannot speak of a center because asymmetric distributions do not have a true center.
[13] In fact, for each positive residual there is a corresponding negative residual.
[14] To better understand the degree of freedom concept the reader may like to consider the following: if we know *n*-1 values of a sample whose size is *n* and we also know the sample mean, then the remaining sample value is unequivocally determined and is not free to take any value. In this case the degree of freedom is *n*-1.
[15] In the next section we shall see that the mean and the variance calculated from the sample are also called *sample mean* and *sample variance.*

The variance does not have the same unit of measurement as the measured variable x, but it does have its square. For this reason it is more suitable to calculate its square root. The *standard deviation* is defined as:

Standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2.3}$$

and is the square root of the variance. The standard deviation, often abbreviated as "SD", has the same unit of measurement as x and it is a direct estimation of the population distribution width, indicated with σ. The standard deviation is the best known spread measure and it is commonly associated with the mean as *mean* ± *SD*.

The standard deviation always has positive values and it is a very good indicator of the width of the symmetric distributions. As stated with regard to the difference between mean and median, when we are dealing with asymmetric distributions, we need more useful spread measures than the standard deviation. The curves shown in Figure 2.5 will help to clarify this point. In graph A the curve is asymmetric on its right side. Since the standard deviation is calculated on the entire data sample, it is significantly influenced by the extreme values. On the other hand, although the curve width in graph B is equally divided between the right and the left side of the mean, this is not the case in graph A. So when we are dealing with asymmetric distributions we have to consider *quartiles* rather than standard deviation. We shall now provide a rigorous definition of quartiles followed by a clarifying example.

Quartiles and percentiles

Let *n* be the size of the sample and let the data be reorganized in ascending order. The *1st quartile* (or, alternatively, *25th percentile*) is the value below which we find the first quarter (*n*/4, 25%) of the observations; similarly, the *2nd quartile* (*50th percentile*) is the value below which we find half (*n*/2, 50%) of the observations; the *3rd quartile* (*75th percentile*) is the value below which we find 75% (3*n*/4) of the observations. The reader will have recognized that the 50th percentile coincides with the median.

50th percentile coincides with median

Let us consider the following example:

21, 36, 4, 85, 4, 56, 87, 65, 12, 24, 2, 54, 9, 32, 30, 26

which indicate the age of $n = 16$ patients. The first step is to rewrite the data in ascending order, as follows:

2, 4, 4, 9, 12, 21, 24, 26, 30, 32, 36, 54, 56, 65, 85, 87

Since *n* is an even number, the 50th percentile (median) is the arithmetic mean of the two central values (26 and 30), i.e. 28. Let us look at the first half of the sample. Since *n*/2 has again an even value, the 25th percentile is the mean of the two central values 9 and 12, i.e. 10.5. In the second half of the sample we calculate the mean of 54 and 56, i.e. 55, which represents the 75th percentile.

2.4. Standard Error of the Mean

In the previous section we introduced the two main parameters for describing the hallmarks of a sample obtained measuring a continuous variable on *n* individuals randomly extracted from the population: mean and SD.

Difference between estimator and estimation

Both the mean and standard deviation of a sample only provide an *estimation* of the true values of mean and SD. The two mathematical relations (2.2) and (2.3) only represent the way by which these estimations are calculated. In order to distinguish the formulas from the calculated values the term *estimator* is often used. In practice, the two estimators mean and SD provide an estimation (m and s, respectively) of the true values of the population (μ and σ, respectively).

In the past many other estimators were proposed as central tendency and data spread measurements and, to be perfectly honest, our preference for mean and SD has not been justified. Without going into details, we may simply state that *mean and SD are the only estimators which have all the features that an estimator must have*.

Since they are only estimations, the numerical values of mean and SD are imprecise evaluations. Their values depend on the sample considered; if, for example, we extract a different sample from the same population and recalculate the mean and SD of this new sample, we will clearly obtain different values. In order to stress this link with the sample, the mean and the SD calculated in (2.2) and (2.3) are also called *sample* mean and *sample* SD.

Sample mean and sample SD

> **Example 2.2. Sample mean and sample SD.** Let us consider the population of all women with breast cancer and let us evaluate the mean size of the tumor. Using the same instruments and diagnostic technique, two different hospitals measure the mean diameter on two different samples made up of 100 patients. The findings of the first hospital are: m = 2.3 cm and s = 1.1 cm; the findings of the second hospital are: m = 2.5 cm and s = 1.0 cm. The results obtained by the two hospitals represent two sample estimations of the true mean and SD.

Example 2.2 demonstrates that even under the same conditions the sample estimations depend on the particular extracted sample. At the same time, in the medical literature we often find many studies reporting several different values for the same variable. If the population from which the samples are extracted is the same and if there are no errors in designing and performing the study, this is simply the result of random sampling[16].

The arithmetic sample mean is the best estimation of the population true value

Although we have not demonstrated it, the arithmetic mean of a sample is the *best* estimation that we have for the true mean value of the measured variable. Let us now ask the following question: "*To what extent* is the sample mean a *good* estimation of the population mean?" In order to answer this question we

[16] The reader should also note that the different values of sensitivity, specificity, accuracy and predictive values of a diagnostic technique found in the literature for a given disease simply represent estimations and have the same features of an arithmetic mean.

have to know *the uncertainty associated with our estimation*: the higher this uncertainty, the less precise our estimation becomes, and vice versa.

Sample distributions: the central limit theorem

Now let us consider the following *ideal experiment*. We extract a large number of samples from the same population, all made up of n individuals[17], and we calculate the mean for each sample. We build a histogram on which we report the observed means on the x-axis, instead of the single observations. Based on the *central limit theorem*, this histogram appears as Gaussian with mean μ (i.e. the same mean as the population) and SD equal to:

$$\frac{\sigma}{\sqrt{n}}$$

The SD of this ideal distribution is the ratio between the SD of the population (σ) and the mean square root of the sample size (n). Therefore, the result of the previous ideal experiment clearly gives rise to a new normal distribution (*the distribution of the sample mean*) with the same center as for the population, but the greater the sample size, the lesser its width is. In fact, if the mean of a sample differs significantly from μ, it is true that by calculating the mean of many samples and then the *mean of the means* we obtain a more precise estimation.

When we talk about the distribution of the sample mean we implicitly refer to the result of the ideal experiment just developed, i.e. a distribution for which on the x-axis we pose the mean of one of the extracted samples instead of the single observations. Moreover, the central limit theorem shows that the distribution of the sample mean appears approximately as Gaussian even if the measured data is not a normal variable, and the larger n is, the better this approximation becomes.

Standard error of the mean (SE)

The SD of the distribution of the sample mean is called *standard error of the mean*, often simply called standard error (SE), defined as:

$$SE = \frac{\sigma}{\sqrt{n}}$$

As the reader may note, since the distribution of the sample mean is the result of an ideal experiment, the SE depends on the true SD, which remains an unknown parameter. On the other hand, in practice we only analyze *one* sample. All we can do is estimate the SE by substituting σ with the SD of the single extracted sample, namely:

$$SE = \frac{s}{\sqrt{n}}$$

The standard error is a measure of the precision of the sample mean estimation

Therefore, once we have a sample, the SE is a measure of the uncertainty associated not with the single measurement but with the arithmetic mean since it is the best estimation of the true value of the population: the lower the SE, the higher the precision of the sample mean is, and vice versa.

The question we posed earlier (*To what extent* is the sample mean a *good* estimation of the population mean?) has not obtained a complete answer. Now

[17] One could say "a sample of samples".

we know how to calculate the uncertainty associated with the sample mean, but we do not have a mathematical relationship between the two quantities. What does it mean that the SE of the mean represents the uncertainty associated with the sample mean? In other words, is the arithmetic mean equivalent to the true value of the population? And if not, how much do they differ?

Confidence intervals

The sample mean, m, may differ significantly from the true value, μ. Therefore, we need a mathematical object able to calculate the probability that m does not differ from μ by more than an arbitrary chosen quantity. This approach focuses on the true value and its goal is the calculation of a probability. Apart from the practical difficulties involved, there is a conceptual error based on the impossibility of knowing the true value. The right approach is, in fact, the reverse one, i.e. to fix a probability (*confidence level*) and obtain the interval that contains the true value with that probability. This interval is called *confidence interval*.

Let us come back to example 2.2. The first hospital found a mean tumor size equal to 2.30 cm with a SD equal to 1.10 cm. Since the size of the analyzed sample is $n = 100$ patients, we may calculate the SE of the mean as:

$$SE = \frac{s}{\sqrt{n}} = \frac{1.1}{\sqrt{100}} = 0.11\,\text{cm}$$

As stated above, the best estimation of the mean tumor size of the population is 2.30 cm with an uncertainty equal to 0.11 cm. Now we want to calculate the interval (in terms of tumor size) that, with a given confidence level, contains the true value. The greater the *a priori* fixed probability, the greater is the width of the interval we are seeking. If, to the limit, we would like *to be sure* and calculate the interval containing the true value with a probability of 100%, the result should be *from zero to infinity*, i.e. the set of all the values the variable may assume[18]. In recent decades it has become widely accepted in the literature that the optimal confidence level is equal to 95%, such that in most cases the 95% confidence interval is calculated (95%CI)[19]. Now we will see how to calculate the 95%CI.

We stated that the mean is the best estimation of the true value, so the confidence interval is obtained summing and subtracting a certain quantity Δm to the mean. In this way we obtain the interval boundaries as:

$$95\%CI = m \pm \Delta m$$

We also saw that the uncertainty of the mean is represented by the SE, so Δm is proportional to the SE, that is:

$$\Delta m = t_{95\%} SE$$

from which:

$$95\%CI = m \pm t_{95\%} SE$$

[18] Note that the measured variable is the size of the breast cancer which obviously cannot have negative values.

[19] We will see the reasons (including historical reasons) for this choice in Chapter 3. However, wider (e.g. with 99% confidence level) or narrower (e.g. with 90% confidence level) confidence intervals can also be calculated.

where $t_{95\%}$ is a quantity that has a *Student's t distribution* with n – 1 degree of freedom. A complete description of the Student's t distribution is beyond the aims of this book. Here we simply state that $t_{95\%}$ represents a numerical value easily retrievable from published datasheets [ALTMAN, 1991].

In Example 2.2 the sample size is $n = 100$ and, therefore, the degree of freedom is n – 1 = 99. From the published datasheets we obtain $t_{95\%} = 1.984$. Therefore:

$$95\%\text{CI} = 2.3 \pm 1.984 \cdot 0.11 = [2.08, 2.52] \text{ cm}$$

Therefore, at the confidence level of 95%, the mean breast cancer size of the entire population lies between 2.08 cm and 2.52 cm; at any rate, there remains a 5% probability that the true value is lower than 2.08 cm or higher than 2.52 cm. *This statement represents a bridge between the features of the sample and that of the population*. Confidence intervals will be discussed in more details in Section 2.6.

2.5. Standard Error of the Difference between Two Sample Means

Here we introduce a simple generalization of the standard error of the mean which will be useful in Chapter 4.

In many circumstances encountered in medical research a comparison is made between the means of two independent samples.

Example 2.3. Myocardial delayed enhancement measurement with cardiac MR imaging. We want to evaluate the difference between two contrast agents (CAs) in terms of *delayed enhancement*. For this reason, a sample of 21 post-ischemic patients undergo a cardiac MR examination with inversion recovery turbo-gradient-echo sequence ten minutes after the injection of 0.1 mmol/kg of CA 1. We measure the signal intensity (SI), expressed in arbitrary units (a.u.), in a region of interest placed in the infarcted myocardium. A second sample of 7 post-ischemic patients is studied with the same technique but with 0.1 mmol/kg of CA 2. Data are reported in Table 2.4.

Example 2.3 shows the typical situation in which two independent samples (whose sizes are n_1 and n_2 and which were extracted from two different populations) are *treated* in different ways: with different drugs or CAs, with different imaging modalities, or even with different techniques of the same imaging modality, etc. In these cases, the question is: "If we find differences in the results, is this effect due to the treatment difference or simply to chance?" In Example 2.3, the mean signal intensity is 43.7 a.u. in the sample treated with CA 1 and 20.4 a.u. in the sample treated with CA 2. It is correct to suspect that this difference, even if large, is simply the result of sample diversity. In fact, nobody may exclude that using both the CAs with the same sample one would obtain very similar results (we will discuss this possibility in the next section). The difference between the delayed enhancement obtained with both the CAs is highly significant ($p = 0.0004$) if analyzed with a non-parametric statistical test (Mann-Whitney U test), for which we refer to Chapter 5. Here we use the data of this example to illustrate a fundamental mathematical parameter.

Table 2.4. Signal intensity for the two contrast agents of Example 2.3

Sample 1	SI (a.u.) CA 1	Sample 2	SI (a.u.) CA 2
1	32.8	1	18.8
2	30.6	2	13.0
3	34.2	3	17.8
4	18.2	4	25.8
5	36.0	5	15.8
6	37.6	6	22.4
7	45.4	7	29.0
8	52.4		
9	66.8	m_2	20.4
10	67.8	s_2	5.7
11	23.2	SE_2	2.1
12	33.0		
13	62.0		
14	51.2		
15	72.2		
16	28.6		
17	29.4		
18	46.0		
19	51.8		
20	33.0		
21	65.8		
m_1	43.7		
s_1	16.1		
SE_1	3.5		

SI = signal intensity; CA = contrast agent; a.u. = arbitrary units.

Distribution of the difference between two means

In the previous section we introduced the standard error as the result of an ideal experiment in which one calculates the mean of several independent samples all with the same size, *n*. Now let us slightly modify this experiment by extracting not one sample but couples of samples at a time; the two samples of a couple have to be treated with the two different treatments that we are comparing. We calculate the mean of each sample of the couple and their difference. In this way we may build the *distribution of the difference of the means* that has variance, σ^2, equal to the sum of the two single variances σ^2_1 and σ^2_2. Therefore, it is:

$$SE(\mu_1 - \mu_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $SE(\mu_1 - \mu_2)$ is the standard error of the difference between the means of the two populations. Since the two true variances remain unknown, we substitute σ^2_1 and σ^2_2 with their best estimations s^2_1 and s^2_2, such that:

$$SE(m_1 - m_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $SE(m_1 - m_2)$ is the standard error of the difference between the two sample means m_1 and m_2 (in Section 4.2 we will see that there is another method for calculating standard error). For Example 2.3:

$$SE(43.7-20.4) = \sqrt{\frac{16.1^2}{21} + \frac{5.7^2}{7}} = 4.1 \text{ a.u.}$$

In practice, our focus shifts from the two single means, m_1 and m_2, to their difference $m_1 - m_2 = 43.7 - 20.4 = 23.3$ a.u., which becomes a new variable whose estimation has an uncertainty equal to 4.1 a.u. Similar to what we saw in the previous section, the confidence interval of the difference of the means is:

The confidence interval of the difference between two means

$$95\%CI = (m_1 - m_2) \pm t_{95\%} SE(m_1 - m_2)$$

where $t_{95\%}$ has to be obtained in the datasheet of the Student's *t* distribution with $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degree of freedom. For Example 2.3, $t_{95\%} = 2.056$ and:

$$95\%CI = (43.7 - 20.4) \pm (2.056 \times 4.1) = [14.8, 31.8] \text{ a.u.}$$

that is, with a 95% confidence level, the true difference between the means of the two populations lies between 14.8 a.u. and 31.8 a.u.

2.5.1. Paired Data

A particular case in comparing two sample means is when each statistical unit of the sample undergoes the two treatments; this circumstance introduces the denomination of *paired data*. For the comparison of the two contrast agents in Example 2.3, the radiologist could repeat the diagnostic examination administering both contrast agents (with a time delay of, for example, one day) to the 21 + 7 = 28 patients.

In this case the starting point for obtaining the confidence interval is to calculate the subject-by-subject difference of the measured values. So our focus shifts from each couple of values to their difference.

In Table 2.5 the column of the difference represents a variable with an almost normal distribution centered about the true value of the difference between the two means m_1 and m_2 with m being an estimation of the true value.

The procedure for the calculation of the confidence interval is similar to the previous one. In fact, we have to calculate the SE in this case too and to use the $t_{95\%}$ value as follow:

$$95\%CI = m \pm t_{95\%} \frac{s}{\sqrt{n}}$$

2.6. Confidence Intervals

In the previous sections we introduced the confidence intervals of the mean and of the difference of two sample means. In this section we wish to provide a broader view of the general concept of confidence intervals.

Table 2.5. Comparison of two sample means for paired data

Individual	1st measurement	2nd measurement	Difference
1	a	b	a-b
2	c	d	c-d
...	...	...	...
...	...	...	...
n	y	z	y-z
Mean	m_1	m_2	m
SD	s_1	s_2	s
			SE = $\frac{s}{\sqrt{n}}$

SD = standard deviation; SE = standard error.

Let us start our discussion by stressing an important hallmark of Gaussian distribution. Under conditions of normal distribution with mean equal to μ and standard deviation equal to σ, 95% of the observations lie in the interval:

$$\mu \pm 1.96\sigma \tag{2.4}$$

This result holds for all values of μ and σ (i.e. for each continuous and random variable) as it is based only on the particular mathematical shape of the Gaussian curve. For example, if we measure a continuous variable for a sample of 500 individuals and calculate mean and SD, then, *on average*[20], 95% of them (450) lie within the interval mean ± 1.96SD. We may also state that if we extract another individual from the population, this will have 95% probability of lying within that interval.

The general mathematical relation (2.4) keeps this feature even if we consider the distribution of the sample mean, i.e. the hypothetical distribution we built in Section 2.4 as the result of an ideal experiment. Thanks to the central limit theorem, the distribution of the sample mean is almost normal, its mean (m) coincides with that of the population and its SD is equal to that of the sample (s) divided by the mean square root of the sample size (*n*). Then 95% of the sample means lie in the interval:

$$m \pm 1.96 \frac{s}{\sqrt{n}} = m \pm 1.96\ SE$$

The format of the latter relationship is very similar to that of the CI95% of the mean, that is:

$$95\%CI = m \pm t_{95\%}SE$$

[20] The feature of normal distribution for which 95% of the observations lie in the interval μ ±1.96σ holds rigorously only for the entire population. For a limited sample we have to state that the 95% of the observations lies within this interval, *on average.*

The two mathematical expressions are very close to one another. When the degree of freedom ($n - 1$) is large enough ($n > 100$), the difference between the *Student's t* and Gaussian distributions is very small. In fact, if n = 101 the number of degree of freedom is $n - 1 = 100$ and $t_{95\%} = 1.98$, very close to 1.96. In practice, with small samples ($n < 100$) one should use the $t_{95\%}$ coefficient, rather than the value of 1.96, because the smaller the sample size, the higher the difference between the two coefficients. The reader should note that it is more correct to use the *t* distribution (therefore the $t_{95\%}$ coefficient) than the normal distribution (therefore the 1.96 coefficient), because in the standard error formula the SD of the population (σ) is estimated by the SD of the sample (s).

What we stated with regard to the mean may also be said about the difference between two means. The reader will have noted that in Sections 2.4, 2.5 and 2.5.1 we followed the same procedure. *The 95% confidence interval of every estimation always has the following format*:

General form of a 95% confidence interval

$$95\%CI = \text{estimated value} \pm \text{coefficient}_{95\%} SE_{\text{estimated value}}$$

The $\text{coefficient}_{95\%}$ depends on the case, but it is always retrievable from published datasheets [Gardner and Altman, 1990].

The confidence interval width is a measure of the point estimation precision

A limited sample provides an imprecise sample estimation of the true value of the population and this imprecision is expressed by the width of the confidence interval: the wider these intervals are, the less precise the estimation is, and vice versa. An estimation with a very wide confidence interval casts more than a shadow of doubt on the reliability of the observed value. Let us suppose, for example, we have measured the specificity[21] of a certain diagnostic modality for the detection of a given disease and have obtained the value of 0.75 with a confidence interval equal to [0.57, 0.93]. Although 0.75 is the best estimation we have, the true specificity could be as low as 0.57 or as large as 0.93, a very wide interval indeed. In this case we cannot trust the obtained estimation because the probability of overestimating or underestimating the true specificity is very high.

Confidence intervals do not provide any information on the estimation accuracy

The confidence intervals shift the focus from the variable estimation, also called *point estimation*, to an interval of value considered as compatible with the population. It is important to understand that *confidence intervals depend on the sample size and on the sample variability and do not provide any information on possible errors of design, implementation and statistical analysis of a study.*

2.7. Confidence Interval of a Proportion

Proportion

Each ratio between two numerical values is a *proportion*. Typical examples are the sensitivity and the specificity (for the detection of a given disease), the predictive values, the fraction of the subjects of a sample who have or do not have a given characteristic, etc. On the other hand, this latter definition represents the most general case: the sensitivity, for example, is defined as

[21] The specificity of a diagnostic modality is the ratio between the true negatives and the sum of the true negatives and the false positives (see Chapter 1).

the ratio between the number of individuals who obtained a positive test result and actually had the disease and the number of all the individuals who actually had the disease.

For the proportions we may also say that the numerical value calculated for a limited sample only represents an estimation of the true proportion and for this reason it is somewhat imprecise. The calculation of the confidence interval of a proportion, p, follows the same general rule we saw in the previous section. As in all cases, one needs to obtain the standard error of p, SE(p), and the coefficient$_{95\%}$. Based on an approximated procedure, we may calculate the standard error using the hallmarks of normal distribution and obtain:

The standard error of a proportion

$$SE(p) = \sqrt{\frac{p(1-p)}{n}}$$

Using coefficient$_{95\%}$ = 1.96:

$$95\%CI\,(p) = p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

The larger n is, the better the approximation. However, the latter formula can only be used when p is not too far from 0.5 (50%); in the extreme cases, namely when p is close to 0 (0%) or to 1 (100%), it may provide nonsense results, with confidence intervals that contain negative values or values higher than 1. For example, if in a sample of 15 post-ischemic patients undergoing contrast-enhanced cardiac MR two of them show delayed enhancement, then p = 2/15 = 0.13 and 95%CI(0.13) = [-0.04, 0.30]: with a 95% confidence level, the true proportion could even be -4%, which is clearly impossible. Conversely, if for example p = 0.92, we could obtain an interval like [0.80, 1.04], with the possibility that the true proportion exceeds 100%.

Use of a binomial distribution

With the diagnostic performance indices (sensitivity, specificity, etc.) we very often observe values close to 1. In these cases, a calculation procedure based on *binomial distribution* should be used. The formula for the calculation of the confidence interval using binomial distribution is more complicated and, for this reason, our advice is to always refer to a statistician or use a statistical software package.

References

American College of Radiology (2003) ACR Breast imaging reporting and data system: Breast imaging atlas, 4th edn. Reston, VA

Altman DG (1991) Practical statistics for medical research. London: Chapman &Hall

Gardner MJ, Altman DG (1990) Gli intervalli di confidenza. Rome: Il pensiero scientifico editore

North American Symptomatic Carotid Endarterectomy Trial Collaborators (1991) Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Eng J Med 325:445-453

Siegel S, Castellan NJ Jr (1992) Statistica non parametrica, 2nd edn. Milan, McGraw-Hill

Soliani L (2007) Statistica applicata alla ricerca e alle professioni scientifiche. Manuale di statistica univariata e bivariata. Parma: Uninova-Gruppo Pegaso; 2:61-97 (http://www.dsa.unipr.it/soliani/soliani.html).

UICC (2002) TNM classification of malignant tumours, 6th edn. Geneva: UICC

3

Null Hypothesis, Statistical Significance and Power

When you get new data,
be ready to change your hypothesis.

FRED TUREK

Observation and theory get on best when they are mixed together,
both helping one another in the pursuit of truth.
It is a good rule not to put overmuch confidence in a theory
until it has been confirmed by observation.
I hope I shall not shock the experimental physicists too much
if I add that it is also a good rule not to put overmuch confidence
in the observational results that are put forward
until they have been confirmed by theory.

ARTHUR S. EDDINGTON

The scientific experiment: to test an idea using facts

The strategic purpose of a scientific work is to demonstrate a hypothesis proposed by the authors. The hypothesis arises from their own anecdotic observations or previous scientific works or from papers previously published by other authors. The primary requirement for a scientific study is *an idea we want to verify by means of a series of facts*. The facts could also be those which other authors have already described (e.g. *meta-analysis* – see Chapter 8). Hence, we can say that the only technology we absolutely need is the neuronal circuitry of our brains as evolved primates.

We need a clear explicit hypothesis

This joke highlights that *a scientific work must always arise from a clear explicit hypothesis*. In the classic case, in order to demonstrate that the hypothesis is true, the scientist designs an *ad hoc experiment*. In this sense, we name the hypothesis as *experimental*[1] *hypothesis*. We should note that the hypothesis is almost always derived from previous observations and from the discussion about them, thus confirming or failing to confirm previous well-established knowledge. There is a continuous interaction between practical experience and development of theories with deep philosophical implications [BELLONE, 2006].

Experimental hypothesis

Experience and theory

[1] Note that here the adjective *experimental* is used with reference to the planning and performing of an experiment which supplies new data to be analyzed with defined methods. In other contexts, it has the restricted meaning of *animal or phantom testing*, to be distinguished from research on humans.

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

The logical flow of a scientific report

However, here we must make a clean break. We have to define *a starting point* (the experimental hypothesis, whatever origin it has) and *a goal* to be reached (the demonstration that the experimental hypothesis is true or false). Between these two extremes there are crucial phases such as the design and performance of the experiment, data collection, data analysis (and not only statistical analysis) and discussion. The structure of a scientific report implies the following logical flow: definition of the experimental hypothesis (at the end of the *Introduction*); design and implementation of the experiment (*Materials and methods*); presentation of the results (*Results*); interpretation of the results (*Discussion*). This matter will be expanded in Chapter 10.

Here we will describe the particular arrangement of the logic of scientific demonstration (with particular reference to the biologic and medical field), which has become relatively well-established in the last 50-60 years.

3.1. Null Hypothesis and Principle of Falsification

An apparent paradox

Now we are confronted with a paradox. The scientist who wishes to demonstrate an experimental hypothesis must adopt the reverse hypothesis to the experimental one. This reverse hypothesis is named statistical hypothesis or *null hypothesis* (H_0). Specialized calculations on the data resulting from the experimental work quantify the probability that the null hypothesis is true. These calculations (the technical core of statistics) can have different logical and computational structure, suitable for the particular setting due to the study design, the type of variables under investigation, and other aspects. *This is the crucial problem of choosing the right statistical test.* If the probability (the well-known p value) that the null hypothesis is true is lower than a predefined threshold (usually 5%, frequently presented as a fraction of the unit, therefore 0.05), we reject the null hypothesis. As an indirect consequence, this rejection allows us to accept the reverse hypothesis, i.e. the experimental hypothesis, which we name H_1. This conceptual system and its terminology were introduced by Ronald A. Fisher (1890-1962) in the 1930s.

Null hypothesis (H_0)

Experimental hypothesis (H_1)

Can we really accept the experimental hypothesis as true?

The debate on the "acceptance" of H_1 as a result of the rejection of H_0 is still open. Formally, obtaining a p value lower than 0.05 only states that rejecting the experimental hypothesis is impossible. The overriding opinion is that we can never consider the experimental hypothesis as *demonstrated*, not even indirectly, at least with regard to the meaning we attribute to the word *demonstration* in mathematics. Statistical significance is a long way from the mathematical demonstration of a theorem, such as in the *Elements* by Euclid. According to some authors, a $p < 0.05$ only allows the experimental hypothesis not to be rejected, keeping it available for further experiments. Even if these experiments confirm the significance, the degree of truth of the experimental hypothesis does not increase. Still other authors argue that a series of experiments concordant for statistical significance only *tends towards the demonstration of* H_1, without reaching an ultimate demonstration. There is a subtle difference between these two ways of thinking. At any rate, the demonstration of H_1 is linked to the falsification of H_0.

Principle of falsification

The principle of falsification plays a well-known role in the philosophy of science and is commonly attributed to Karl Popper. Actually, at least in the logical context of statistical thought, it should be attributed to Ronald A. Fisher. Moreover, while in Popper's thinking it is derived from *"simple epistemological affirmations, mainly left to the reader's intuition"*, in Fischer's thinking it is based on *"well-established mathematical-probabilistic models"* [Caracciolo, 1992]. According to Luca Cavalli-Sforza:

Karl Popper or Ronald A. Fisher?

"Even all the recent epistemological work, from Kuhn to Popper, seems to me overdone. The Vienna Circle […] already stated the same things. In the last fifty years all we have had is a great deal of popularization, with these ideas perhaps being presented more clearly, more frequently with synonyms or more popular neologisms (as with the idea that only «falsifiable» theories are scientific" [Cavalli-Sforza L and Cavalli-Sforza F, 2005].

Why do we work against our experimental hypothesis?

What is the reason for this logical process that tests the reverse hypothesis of the one we wish to demonstrate? To answer this question let us consider the classical experimental design aimed at testing whether two samples are different for a defined characteristic. Here the null hypothesis is that the two samples are drawn from the same population and that the observed difference is caused only by random sampling. The hard core of the problem is the variability intrinsic to all biologic phenomena. In fact, if we draw two random samples from the same population and we measure a defined characteristic, the probability of observing a difference is very high. As a consequence, when we observe a difference between two groups or two samples, the first thing we should exclude is that this difference is simply due to the effect of variability within the same population from which the two samples could have been drawn. In other words, the observed difference would not have the meaning that the two samples are different because they were drawn from two populations which actually are different for the feature under investigation. As we will see, this difference due to random sampling has high probability of being *not significant*. This is the reason for which *we reject the null hypothesis when the difference is significant* and *we accept the null hypothesis when the difference is not significant*.

Biological variability (and that related to the measuring process) is the basic problem

Significant and not significant differences

In this reasoning we did not consider the possibility that our data are flawed by some *bias* (i.e. a systematic distortion or error). Bias may generate a significant but false distortion, as with a defect in random sampling or a systematic error in measuring the variable under investigation in one of the two samples we are comparing. This issue will be systematically examined in Chapter 9.

Bias which cannot be eliminated

The reader should now understand the following general concept: sometimes we can correct for bias using statistical tools in data analysis but *commonly there is no way to eliminate the effect of bias in study design or data acquisition*. Only correct study design (which should be carefully planned together with the definition of H_1 and H_0 before starting with data acquisition) enable us to minimize the sources of bias. Only in this way can we pose the crucial question: *Is the observed difference due to a real difference between the two different populations from which the two samples have been drawn or is it due to variability within the single population from which both samples have been drawn?*

The crucial question

3.2. Cutoff for Significance, Type I or α Error and Type II or β Error

How should we define the *cutoff*, i.e. the decisional threshold, we use to decide whether to accept or reject the null hypothesis? This problem is very similar to the problem regarding the distinction between positives and negatives of a diagnostic examination (see Chapter 1). In this case we also have four possibilities:

- *true positive*, when we judge an existing difference as real, i.e. as not attributable to random sampling;
- *true negative*, when we judge a non-existing difference as not real, i.e. as attributable to random sampling;
- *false positive*, when we judge a non-existing difference as real, i.e. as not attributable to random sampling;
- *false negative*, when we judge an existing difference as not real, i.e. as attributable to random sampling.

However, in statistical test applications the false positive case and the false negative case are given a different name:

- the false positive is named *type I error* or α *error*;
- the false negative is named *type II error* or β *error*.

The acceptable level of error (i.e. the cutoff definition) for both type I and type II is represented as a probability.

False positive = type I error = α error

The cutoff for type I error is conventionally fixed at 5%. The null hypothesis is refused when the statistical test tells us that the *probability of obtaining a difference equal to or larger than the observed one is lower than 1 in 20, i.e. 5%* ($\alpha = 0.05$). The cutoff is rarely more restrictive, for instance equal to 1% ($\alpha = 0.01$), or less restrictive, for instance equal to 10% ($\alpha = 0.1$). The *p* value obtained with the statistical test, i.e. the observed level of significance, defines the acceptability of the null hypothesis (H_0): if the cutoff (α) is chosen at 0.05, we consider the *p* values lower than the cutoff ($p < 0.05$) as significant. The smaller the *p* value, the more improbable H_0 becomes, and as a consequence the experimental hypothesis (H_1) is more probably true, within the philosophical limitations examined in Section 3.1.

α = 0.05

$p < 0.05$

The definition of the *probability of obtaining a difference equal to or larger than the observed one* implies the following reasoning: if I repeat the same experiment *n* times randomly drawing two samples of subjects from the same population, how many times do I observe a difference equal to that previously observed or larger due to the combination of the intrinsic variability of the population with the random sampling?

Relation with Gaussian distribution

The careful reader will have noticed that even though the cutoff level is conventional and sometimes may be increased or reduced, the choice of 5% ($\alpha = 0.05$) can be related to a basic feature of Gaussian or "normal" distribution (see Section 2.4) whereby 95% of data are in the range mean ± 1.96 standard deviation. Due to the bell shape of normal distribution, this 5% is evenly distributed between the two tails of the curve, as shown in Figure 3.1.

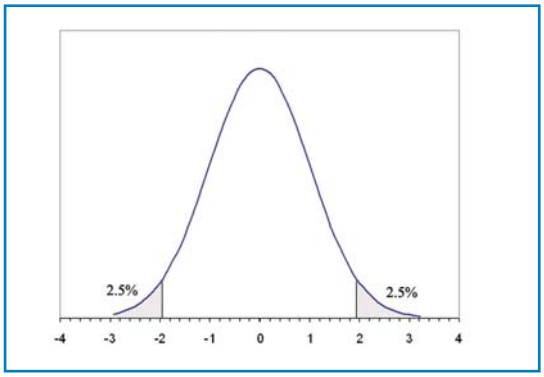


Figure 3.1. The two tails of normal standard distribution. The graph shows two tails at the extremes of the distribution, positioned at the values of z = ± 1.96, each accounting for 2.5% of the statistical units of the population (total equal to 5%, i.e. 0.05).

One-tailed, two-tailed

The result of a *two-tailed statistical test* takes into consideration the possibility that the two compared measurements of the same variable (e.g. *a* and *b*) can be significantly different for either $a > b$ or $a < b$, i.e. it takes into consideration both tails of the distribution (2.5% for $a >$ b and 2.5% for $a < b$). If the test driver can a priori exclude one of the two possibilities, one of the two tails of the distribution can be ignored (*one-tailed statistical test*). As a consequence, the error probability is halved and the significance of the test result is doubled. The same result which gives $p = 0.9$ (not significant) using a two-tailed test, gives $p = 0.045$ (significant) using a one-tailed test, since the cutoff remains unchanged ($\alpha = 0.05$). However, if the test driver is not absolutely certain that the difference between the data may occur in only one direction, the use of a two-tailed test is recommended.

False negative = type II error = β error

β = 0.20 or β = 0.10

The cutoff for type II error should be commonly chosen equal to 80% or 90%. This means that we accept to not consider an existing difference as real no more than one in five times (80% cutoff) or no more than one in ten times (90% cutoff). In the first case the β error is 0.20, in the second case the β error is 0.10. However, the definition of the β error of a study is much less common in the (radiologic) literature than the definition of the α error, as we shall see in the next section and in Chapter 8 (Section 8.8).

3.3. Statistical Power

In published articles, the explicit declaration of the cutoff is very frequent for α error (almost always $\alpha = 0.05$) and much less common for β error. The reason is that most published articles report results with at least one statistically significant difference. In these articles, the possibility of a β error is excluded by the detection of one or multiple significances with $p < 0.05$. In other words, if we have rejected the null hypothesis and *not rejected* (indirectly accepted) the experimental hypothesis, this implies that the statistical test has given a positive result with increasing probability that this is true, the smaller is the *p* value (*p* is the residual probability of false positive). In all these cases, given the positive result, there is no sense to questioning the probability of false negative (β) and true negative (1 – β). *If the result is positive, it cannot be negative.*

p as the residual probability of false positive

The β error problem

The β error problem arises when we do not obtain any significance ($p \geq 0.05$). In this case the null hypothesis is accepted and the experimental hypothesis rejected. Here the question is: which β error (i.e. type II error, false negative) was considered acceptable in the study? In other words, did the study have sufficient *power* to detect a difference judged clinically relevant as significant? If β is the probability of a II type error, the *power* is the complement to 1 of β:

Power

$$\text{power} = 1 - \beta$$

Similarity between statistical testing and clinical diagnosing

We have already mentioned the useful similarity between true and false positives and negatives of a statistical test and those of a diagnostic examination. In diagnostics, these numbers give rise to performance quantification in terms of sensitivity, specificity etc. Even though there is a logical parallelism between diagnostic sensitivity and statistical power (1 – β) just as there is between diagnostic specificity and statistical complement to 1 of the α error (1 – α), the diagnostic terminology classically does not apply to statistical tests. In Table 3.1 we present the comparison between the 2 × 2 contingency table of a diagnostic examination and that of a statistical test. Given the definitions of α error and β error, the true positives become 1 – β, i.e. a portion of the unit, equal to sensitivity. Similarly, the true negatives become 1 – α, i.e. a portion of the unit, equal to specificity.

Four factors determing the power

What does power depend on? Basically, on four factors:

1. On the α *error* chosen by the authors. Larger α, less probable the acceptance of the null hypothesis and lower the risk of type II error; smaller α, more probable the acceptance of the null hypothesis and higher the risk of type II error;
2. On the spread of the observed values, i.e. on the *variability of the phenomenon* under investigation. When two samples are compared, this variability is added to the effect of random sampling: the lower the variabil-

Table 3.1. Comparison between the 2 × 2 contingency table of a diagnostic examination (A) and a statistical test (B)

A		Truth	
		Disease present	Disease absent
Diagnostic examination	Positive	True positives (TP)	False positives (FP)
	Negative	False negatives (FN)	True negatives (TN)
B		**Truth**	
		H_0 false; H_1 true	H_0 true; H_1 false
Statistical test	Positive ($p < 0.05$)	(1 – β)	Error α
	Negative ($p \geq 0.05$)	Error β	(1 – α)

ity, the lower the possibility that the two means of two samples drawn from two different population are similar, causing a type II error; the larger the variability, the larger the probability of a type II error. In fact, the numerator of the standard error of the sample mean is the standard deviation, i.e. a parameter measuring the spread of the observed values (see Chapter 2);

3. On the *amount of the minimal difference judged as clinically useful to demonstrate*. The larger this minimal difference, the smaller the probability of type II error. This is simply due to the fact that detecting large differences is easier than detecting small differences. Moreover, although real but undetected, small differences are not a real type II error in medicine, since we have a priori considered them as clinically irrelevant;
4. On the *sample size*. The larger the samples, the more frequently their means tend to be the same as the single population from which they could be drawn. As a consequence, the probability of detecting small real differences increases.

Sample size: the only factor we can handle to increase the power

Now let us consider that: (1) α is almost always chosen equal to 0.05; (2) the variability of the phenomenon under investigation cannot be substantially changed for the defined clinical and technical setting; (3) the amount of minimal difference judged as clinically useful to demonstrate depends on clinical considerations external to the study itself (a kind of precondition of the study, as with pathophysiologic knowledge derived from previous studies). Hence, *the only factor we can handle to increase the power of a study (i.e. to reduce the probability of type II error) is the sample size*. When we design a study, we should define not only the level of α error, but also the amount of minimal difference judged as clinically useful to demonstrate, the sample size, and the power of the study ($1 - \beta$). Remember that the power of the study is basically determined by the sample size (see Chapter 8).

Diagnostic versus statistical reasoning

The similarity between the results of a diagnostic examination and the results of a statistical test deserves one more comment. *In the two fields the logical path is inverted.*

In diagnostics we put sensitivity in the front row, specificity in the second row. In fact, diagnostic reasoning arose from clinical activity on symptomatic subjects (the patients). In this setting the detection of an existing disease and avoidance of false negatives is the main task. Only recently have we begun to screen asymptomatic subjects where the first priority is to avoid false positives (otherwise we would medicalize the normal population – see Section 1.3). *In diagnostic reasoning, sensitivity (β error) comes first.*

Conversely, statistical testing was introduced in medicine due to the need to judge the efficacy of new treatments. In this setting the main aim is to avoid falsely judging a new therapy as better than the placebo or standard of care, i.e. we must avoid false positives. The calculation of the study power to quantify (and minimize as much as possible) the risk of false negative (to judge an effective therapy as noneffective) is on a second line of reasoning. *In medical statistical testing, α error (specificity) comes first.* Thus, we have an inverted logical path in comparison with diagnostic reasoning. Statistical reasoning follows alpha… betic order!

3.4. Why 0.05?

Do we distinguish false from true or improbable from probable?

This is not a contrived question. We cannot simply answer the question by stating that from around the 1960s onwards scientists increasingly chose $\alpha = 0.05$ as an established convention. In fact, this cutoff seems to have the magic ability of distinguishing truth from untruth. This appears not very "scientific".

Firstly, *a statistical cutoff separates what is probable from what is improbable as regards the null hypothesis, not untruth from truth*. The philosophical difference between the quantification of the uncertainty and the ultimate demonstration of the experimental hypothesis was discussed above. However, even if we remain in the field of probability, another question needs to be asked: Why accept the null hypothesis with $p \geq 0.05$ and reject it with $p < 0.05$? In other words, why does the scientific community universally accept that $p < 0.05$ implies statistical significance?

$p < 0.05$: historical and methodologic reasons

Historical and methodologic reasons explain this fact [Soliani, 2007]. In the early part of the last century, books on statistics reported many tables with long series of p values. Ronald A. Fisher (1890-1962) shortened the tables previously published by Karl Pearson (1857-1936), not only for reasons of editorial space but probably also for copyright reasons (Fisher and Pearson were not on good terms). Some p values were selected and became more important. This was due to the fact that Fisher wrote for researchers (the users) and not for expert in statistics (the theoreticians). According to Soliani, Fisher *"provides a selection of probabilities which simplifies the choice and helps in decision making"* [Soliani, 2007]. Fisher himself attributed a special status to $p = 0.05$, asserting explicitly: *"The value for which p = 0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not"* [Fisher, 1956].

$p < 0.05$ according to R.A. Fisher

However, Fisher and his school (including Frank Yates, 1902-1994) were not unshakeable for the use of the 0.05 cutoff. On numerous occasions, they proposed a soft and problematic interpretation [Soliani, 2007], taking into account factors of uncertainty, first of all the sample size. If n is small, the interpretation of p values near the cutoff is uncertain.

$p < 0.05$ according to J. Neyman and E.S. Pearson

From the late 1920s/early1930s, Jerzy Neyman (1894-1981) and Egon S. Pearson (1896-1980), son of Karl, proposed a different approach – *hypothesis testing*. In this conceptual framework, the cutoff value for p should be defined before the experiment and the statistical result is taken into consideration only as under the cutoff (significant) or equal to or over the cutoff (not significant). The real value of p is barely relevant. Fisher was against this attribution of an absolute value to the predefined cutoff and highlighted the need to report in manuscripts the exact p value and to interpret its evidence. This conflict of opinion can also be related to the debate between the frequentist statisticians (like Fisher and Yates) and Bayesian statisticians (like Neyman and Pearson) [Soliani, 2007].

Decision-making

The Neyman-Pearson approach is useful in *decision making*, but it has evident limitations when we have small samples, especially when categorical variables make the use of non parametric tests mandatory. In these cases, changing only one result may modify the p from values near to 0.01 to values over 0.05. With large samples and asymptotic distributions we have more certainty.

Moreover, modern computers permit the calculation of exact p values which can be presented to the reader for an evaluation of the level of evidence.

Always report the p values

What is the current practice in medical and radiologic journals? The α error is almost always defined equal to 0.05. Hence, values of $p < 0.05$ are considered significant and values of $p \geq 0.05$ are considered not significant. The use of a different cutoff (e.g. 0.1 or 0.01) should be explicitly justified (a job for a professional statistician). *It is recommended that exact p values always be presented*, at least for the values which imply a significance (< 0.05), so that the reviewer and the possible reader of the journal can evaluate the amount of uncertainty associated with your p. Many journals accept that for $p \geq 0.05$ only the non significance (*n.s.*) is reported. However, exact p values are increasingly reported also for $p \geq 0.05$.

Therefore, we are in an intermediate situation between the rigid use of the cutoff and a more debated evaluation of the p value we obtained. At any rate, as we will see in the next section, even a rigid interpretation cannot ignore the difference between statistical significance and clinical significance.

3.5. How to Read a *p* Value

The medical research conducted in recent decades has been characterized by the application of this conceptual system (hypothesis H_0 and hypothesis H_1, statistical significance) and of technical statistical tools (parametric and non-parametric tests – see Chapters 4 and 5). Today, an *original article* (see Chapter 10) without at least a minimal statistical analysis is barely acceptable by a peer-reviewed journal. Moreover, original articles reporting one or multiple statistical significances ($p < 0.05$) and thus demonstrating the efficacy of new diagnostic or therapeutic procedures have a higher probability of being published than articles reporting non significant results ($p \geq 0.05$)[2]. This implies a selection in publishing medical researches known as *publication bias*.

How then should a p value be interpreted?

Evaluate the real value of p

The first rule is to evaluate its *real amount*. Knowing that "$p < 0.05$" is not enough. There is a huge difference between $p = 0.049$ and $p = 0.0049$: the probability of being in error when stating there is a real difference between the two samples changes from nearly 1 in 20 to nearly 1 in 200. We recommend always giving the exact p value, with at least three decimals. This practice is increasing even for p values ≥ 0.05, which for a long time have simply been reported as not significant (n.s.).

The p values directly measures the probability of a false positive result of the statistical test

Remember that *the p value directly measures the probability of a false positive result of the test*, i.e. the probability of rejecting H_0 when H_0 is true and as a consequence of accepting H_1 when H_1 is false. For example, suppose we compare the sensitivity for a given disease of a new advanced imaging technique (New) with that of the old technique (Old). If we obtain a $p < 0.05$ in favor of a higher sensitivity of New compared with that of Old, the smaller the p value is, the lower the error probability is which affirms that New is more sensitive

[2] In this book we do not examine the non-inferiority studies, for which we recommend the consultation of specialized texts.

than Old. *It is counterintuitive that the amount of p does not measures the amount of the difference in sensitivity between New and Old; p only measures the reliability of our affirmation that New is more sensitive than Old, not the extent to which New is more sensitive than Old.*

Look at the raw data!

A simple recommendation is to look at the data, the real raw numbers given as results before any calculation or processing. For the example proposed above, this involves asking how many true positives does New have compared to Old. *Compare the two sensitivities by evaluating the two ratios which generate them.*

Example 3.1. Comparative study of the sensitivity of the New technique versus the Old technique for disease X. Out of 1,000 patients, at the reference standard 682 are found to be affected and 318 not affected with X. The sensitivity of Old is 0.72 (490/682), whereas the sensitivity of New is 0.73 (498/682). In fact, New detects all the 490 true positive at Old plus another 8 which are false negative at Old. The sensitivity increases by about 1%, from 72% to 73%, with $p = 0.008$ (McNemar test – see Chapter 5), i.e. lower than 0.01, and therefore with a high statistical significance. Thus, we have less than 1% probability of being in error when stating that New is more sensitive than Old for the disease X. *However, the real amount of the gain in sensitivity (only 1%) is clinically not relevant.*

The *p* values does not quantify the amount of the difference between the two samples

Always bear in mind that *p does not quantify the amount of the difference between two samples,* rather *p* values quantify the reliability of our rejection of H_0. To see its practical meaning, we must look at the raw data. Use your *common sense* to evaluate the real difference, even if this is statistically highly significant.

References

Bellone E (2006) L'origine delle teorie. Turin: Codice Edizioni

Caracciolo E (1992) Introduction to the 2nd italian edn of: Siegel S, Castellan NJ Jr. Statistica non parametrica: Milan. Mc-Graw-Hill

Cavalli-Sforza L, Cavalli-Sforza F (2005) Perché la scienza. L'avventura di un ricercatore. Milan: A. Mondatori, 338

Fisher RA (1956) Statistical methods for research workers. New York: Hafner, 44

Siegel S, Castellan NJ Jr (1992) Statistica non parametrica. Italian edn. edited by Caracciolo E, Milan: McGraw-Hill, 14

Soliani L (2007) Statistica applicata alla ricerca e alle professioni scientifiche. Manuale di statistica univariata e bivariata. Parma: Uninova-Gruppo Pegaso, Ch 4:8-11. (http://www.dsa.unipr.it/soliani/soliani.html)

4

Parametric Statistics

Only naughty brewers deal in small samples.

KARL PEARSON

The importance of Gaussian distribution

In Chapter 2 we introduced the fundamental hallmarks of Gaussian distribution, thereby neglecting many other theoretical distributions which may also be found in medical research. This preference is based on the simple fact that, even with some limitations, almost all the other distributions tend to coincide with normal distribution. The links between the theoretical distributions allow one to analyze the sample data using, at the first level of approximation, statistical techniques based on the hallmarks of Gaussian distribution. When, for example, we use the coefficient 1.96 for the confidence interval calculation, we are implicitly using a well known hallmark of normal distribution. If, on the other hand, we want to be rigorous, we have to use the correct theoretical distribution, case by case.

Statistics founding fathers

The foundations of Statistics were mainly laid by Lambert A.J. Quetelet (1796-1874), Francis Galton (1822-1911), Karl Pearson (1857-1936), William S. Gossett (1876-1937), Ronald A. Fisher (1890-1962) and George W. Snedecor (1881-1974). As stated in previous chapters, one of the goals of Statistics is to *infer* the results observed in a limited sample to the entire population. However, this approach was born around 1925, about 20 years after the publication of research by William Sealy Gossett, in the journal *Biometrika*, conducted on samples of Guinness beer, the company he was working for due to the lack of an academic job [SOLIANI, 2007]. So as not to reveal trade secrets to rival breweries, Gossett's employment contract restricted him from publishing the results of his research. To circumvent this problem, he therefore published his results using the pseudonym "A. Student"[1]. These studies were published between 1907 and 1908.

[1] In a controversial fashion about the academic world.

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

The dispute which gave rise to modern statistics

Before the publication of Gossett's studies, statisticians were focused on the exploration of theoretical distributions, namely the distribution of the entire population[2]. Karl Pearson responded to the thesis of A. Student stating: "*Only naughty brewers deal in small samples*". Later, Ronald A. Fisher took up the defense of Gossett replying:

"*… the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data*" [quoted in SOLIANI, 2007].

We wanted to report this debate not only to provide an idea about the historical reasons that very often lie behind universally accepted theories, but also to comment on the birth of *modern statistics* or *practical statistics*, which deals with the methods suitable for analyzing small samples. Before this time, nobody took into account the question of checking whether two samples belonged to the same or to different populations, i.e. whether the two samples were different for a variable, an effect, a treatment, etc. Before practical statistics was born, the differences between two or more populations were studied (when possible) by comparing the corresponding theoretical distributions.

Differences between theoretical and modern statistics

In practical statistics, one of the most frequent practices is the comparison between two or more samples initially considered as belonging to different populations. The typical example is that in which a group of individuals is treated with the standard treatment and a second group with the experimental treatment. In this case the logical approach is the following. One hypothesizes that the first group belongs to the population treated in the standard way and that the second group belongs to the population treated in the experimental way. If the diversity between the two treatments produces a real statistically significant effect, then they actually are different populations. On the other hand, if the two treatments do not produce statistically significant differences, then the two populations coincide with each other and the two samples have both been extracted from the same population.

In this chapter the main parametric statistical tests are presented. Student's *t* test will be given the broadest treatment because it is very easy to handle. Moreover, it allows us to discuss the general approach which is adopted with all parametric statistical tests.

4.1. The Foundations of Parametric Statistics

Gaussian distribution is characterized by only two *parameters*: mean and standard deviation. Once we know these parameters, the distribution is unequivocally defined. We saw in Chapter 2 how to obtain an estimation

[2] In theoretical statistics the population represents an infinitely large sample of statistical units, not necessarily made up of human beings.

Table 4.1. Necessary conditions for applying parametric statistical tests

Object	Description
Type of variables	Continuous or at least interval variables
Distribution of the variables	Normal or near-normal distribution
Variances	Variances equal to each other (*homoschedasticity*)

of mean and standard deviation based on the sample data. The set of analysis techniques whose logical approach is based on the features of normal distribution make up *parametric statistics*. Alongside parametric statistics is *non-parametric statistics* which does not rely on the features of normal distribution.

Using parametric statistics requires that some hypothesis be verified

Parametric statistics provides very powerful methods of analysis, but their application requires that some hypotheses be verified, hypotheses which are rarely encountered in radiologic research. A list of the necessary assumptions for using parametric statistics is given in Table 4.1.

Requisition for continuous variables

It is clear that parametric statistical tests can only be applied in the comparison of continuous variables or variables measured with interval scales. However, the classification of a radiologic examination is very often an ordinal or dichotomous (positive/negative) result. The typical example of an ordinal scale of measurement is the BI-RADS® [American College of Radiology, 2003] for mammography: 0, inconclusive; 1, negative; 2, benign; 3, probably benign; 4, suspicious abnormality; 5, highly suggestive of malignancy; 6, already known malignancy. If we consider, for example, two samples differing in the diagnostic examination used for their detection and for which the measured variable is the BI-RADS® score, they cannot be compared using parametric statistical tests. We shall see in the next chapter that the statistical analysis of categorical data always requires non-parametric methods.

Requirement for normal distribution

The second condition for applying parametric statistics involves the shape of the distribution of the measured variable. In order to apply parametric statistical methods we always have to verify that the sample data have normal distribution or, at least, provide some reasons for explicitly supposing this to be the case. The use of parametrical methods with non-normally distributed samples may provide false significant results. The further the data distribution is from the Gaussian curve, the greater the error we make.

Requirement for homoschedasticity

The third condition for using parametric methods, which is almost always not verified, is *homoschedasticity*. This term indicates the situation in which the compared variable has the same variance in the two populations. In practice we analyze the possible difference between, for example, two sample means, even though our hypothesis is that the populations from which the samples are extracted have the same variance. Let us suppose we extract two random samples of breast cancers in symptomatic women (clinical mammography) and asymptomatic women (screening mammography) and we assess the difference of the mean tumor diameter. Even if we suspect that the mean diameter is larger for clinical mammography than for screening mammography, to use a parametric test we have to suppose (or to demonstrate)

that the variance is the same for the two samples, a condition which is not necessarily true.

Homoschedasticity hypothesis simplifies the theory

The characteristic of homoschedasticity is not an intuitive concept. The main reason for taking this condition into account is the presence of the true variances at the numerator and at the denominator of the mathematical formula developed in parametric methods. Although the true values are never known, if they coincide with each other they disappear from the ratio, so leaving the formula independent of them.

The reader should be aware that the concept of homoschedasticity may give rise to confusion. Although this feature is among the necessary assumptions for applying parametric methods, it is possible to modify the theory of Student's *t* test so as to include the most general case of non-homoschedasticity (*heteroschedasticity*). The inclusion of the general case is not intended to make the discussion more difficult to understand, but it is necessary so that the reader may understand the results provided by statistical software packages. In fact, when performing Student's *t* test, these computer programs calculate the *p* value both with and without the hypothesis of homoschedasticity.

Lastly, radiologic studies often deal with very small sample sizes which makes checking the hypotheses reported in Table 4.1 all the more difficult. Therefore, most of the time radiologists will prefer non-parametric statistical methods.

4.2. Comparison between Two Sample Means: Student's *t* Test

In Chapter 3 we discussed the logical approach which lies behind the *statistical tests* for the verification of the null hypothesis H_0. There we stated that if the probability of obtaining a result equal to or even larger than the observed one (probability which is calculated with the null hypothesis being true) is lower than the threshold value, conventionally chosen as 5%, then the null hypothesis has to be rejected. Now let us consider how to calculate this probability when comparing two sample means.

We retrieve the definition of the 95% confidence interval of a sample mean m:

$$95\%CI = m \pm t_{95\%}SE$$

where SE is the standard error of the mean, equal to the ratio between the sample standard deviation (s) from which m is calculated and the mean square root of the sample size. Therefore, once we have a statistical sample, the width of the 95%CI depends (other than on m and on s) on the $t_{95\%}$ coefficient, which is provided by suitable tables [Altman, 1991].

By definition, 95%CI contains the true value of the population (also called *expected value*) with a probability equal to 95% and we *hope* that the width of this interval is as small as possible. As the width reduces we gradually have a more precise estimation of the expected value; when to the limit, this width becomes zero, the 95%CI coincides with the expected value. Without going into the mathematical details, we may rewrite the previous equation as follows:

$$\text{expected value} = \text{observed value} - t_{95\%}SE$$

where the expected value takes the place of the 95%CI, while the observed value is only an alternative way of indicating the sample mean m.

From the last equation we have:

$$t_{95\%}SE = \text{observed value} - \text{expected value}$$

from which:

$$t_{95\%} = \frac{\text{observed value} - \text{expected value}}{SE}$$

Conceptual bridge between confidence intervals and statistical hypotheisis verification

The aim of this mathematical process is not to propose a new way of calculating confidence intervals; indeed, in the last equation the expected value remains unknown. However, its utility becomes clear when we want to compare two sample means, m_1 and m_2. In this case we have a statistical test whose null hypothesis is that the two means are not significantly different from each other.

When comparing two sample means we focus on the difference $(m_1 - m_2)$ which, if the null hypothesis H_0: $m_1 = m_2$ is true, produces an expected value equal to zero. Now we have all we need to calculate $t_{95\%}$, as:

$$t_{95\%} = \frac{(m_1 - m_2) - 0}{SE(m_1 - m_2)}$$

where $SE(m_1 - m_2)$ is the standard error of the difference of the two sample means, whose calculation is illustrated in Sections 2.5 and 2.5.1. The $t_{95\%}$ value has to be compared with the same tables used for the calculation of the confidence intervals [Altman, 1991], from which one obtains the probability *p* which allows us to establish whether the difference $(m_1 - m_2)$ is statistically significant. From a mathematical point of view, $t_{95\%}$ can take positive or negative values[3] and the larger its absolute value, the lower the corresponding *p* value is and, therefore, the higher the significance of the difference between m_1 and m_2 is. Conversely, the closer $t_{95\%}$ is to zero, the larger the corresponding *p* value is and the lower the significance of the difference is.

The larger the $t_{95\%}$ value, the lower the correspondent *p*-value

The theory reported here was developed by Gossett and the statistical test performed by the calculation of the previous equation is known as *Student's t test* for the comparison of two sample means.

Student's *t* test may apply both to independent and paired data

As we stated in the Sections 2.5 and 2.5.1, in practice we may encounter two circumstances: the case of paired data and the case of independent data. In the first case, the two compared statistical samples are obtained by measuring the same continuous variable for a group of individuals before and after a certain treatment, where the term *treatment*, as usual, has to be interpreted in the most general way. The same is when two different treatments are applied to the sample. In the case of independent data, the two sample means rely on different samples, namely on

[3] The reader should note that since Student's *t* distribution is symmetric about the zero, in the corresponding published tables only the positive values are reported, with which the absolute value of $t_{95\%}$ should be compared.

two groups made up of different individuals. *At any rate, the logical approach of Student's t test is the same for the two cases and the sole difference is the calculation of the standard error of the two mean difference, SE($m_1 - m_2$).*

Let us consider the following example.

Example 4.1. Measuring myocardial delayed enhancement in cardiac MR imaging. Let us suppose we want to evaluate the difference in delayed enhancement of the myocardium provided by two contrast agents (CAs). A sample of 50 post-ischemic patients undergo a cardiac MR with inversion recovery turbo-gradient-echo sequence ten minutes after the injection of 0.1 mmol/kg of CA 1. The signal intensity (SI), expressed in arbitrary units (a.u.), is measured in a region of interest placed in the infarcted myocardium. A second sample made up of another 50 post-ischemic patients is studied with the same technique but using 0.1 mmol/kg of CA 2. Data are reported in Tables 4.2 and 4.3.

Now the question is: "Is the observed difference between the means 50.7 – 39.0 = 11.7 a.u. statistically significant? Or is this difference due to chance?" In other words: "Should we accept or reject the alternative hypothesis H_1: m_1 = 39.0 a.u. ≠ m_2 = 50.7 a.u.?" What we are proposing is a typical comparison between two sample means for independent data in which $n_1 = n_2 = 50$.

Verifying the conditions for applying the *t* test

The signal intensity is a continuos variable. Therefore, to apply Student's *t* test we have to verify that the data are normally distributed and that the variances of the two samples are approximately equal to each other. Figures 4.1 and 4.2 show the histograms of the signal intensity of Example 4.1. Since the graphs have a near Gaussian shape, we are quite sure that signal intensity is a random variable in both samples; moreover the two curves have about the same width. The three assumptions reported in Table 4.1 for the application of the parametric tests have all been verified.

Table 4.2. Signal intensity measurements after the administration of CA 1

Individual	SI (a.u.)	Individual	SI (a.u.)	Individual	SI (a.u.)
1	38.74	19	39.39	37	42.25
2	39.26	20	40.30	38	36.40
3	39.13	21	39.65	39	36.50
4	40.56	22	38.48	40	35.62
5	37.18	23	41.99	41	39.52
6	38.61	24	36.27	42	39.65
7	37.40	25	37.05	43	40.30
8	40.17	26	37.57	44	38.48
9	40.56	27	40.82	45	38.74
10	38.22	28	41.08	46	38.60
11	37.96	29	39.13	47	39.00
12	38.87	30	39.78	48	39.13
13	38.30	31	39.91	49	38.74
14	37.18	32	38.61	50	39.13
15	41.34	33	38.87		
16	41.86	34	38.09	m_1	39.0
17	39.26	35	39.13	s_1	1.5
18	38.87	36	39.26	SE_1	0.2

SI = signal intensity; CA = contrast agent; a.u. = arbitrary units.

Table 4.3. Signal intensity measurements after the administration of CA 2

Individual	SI (a.u.)	Individual	SI (a.u.)	Individual	SI (a.u.)
1	50.36	19	51.21	37	54.93
2	51.04	20	52.39	38	47.32
3	50.87	21	51.55	39	47.45
4	52.73	22	50.02	40	46.31
5	48.33	23	54.59	41	51.38
6	50.19	24	47.15	42	51.55
7	48.62	25	48.17	43	52.39
8	52.22	26	48.84	44	50.02
9	52.73	27	53.07	45	50.36
10	49.69	28	53.40	46	50.18
11	49.35	29	50.87	47	50.70
12	50.53	30	51.71	48	50.87
13	49.79	31	51.88	49	50.36
14	48.33	32	50.19	50	50.87
15	53.74	33	50.53		
16	54.42	34	49.52	m_2	50.7
17	51.04	35	50.87	s_2	1.9
18	50.53	36	51.04	SE_2	0.3

SI = signal intensity; CA = contrast agent; a.u. = arbitrary units.

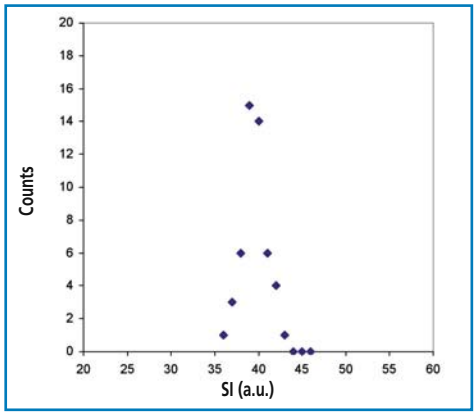


Figure 4.1. Histogram of the signal intensity (SI) measured in arbitrary units (a.u.) for the data of Table 4.2.

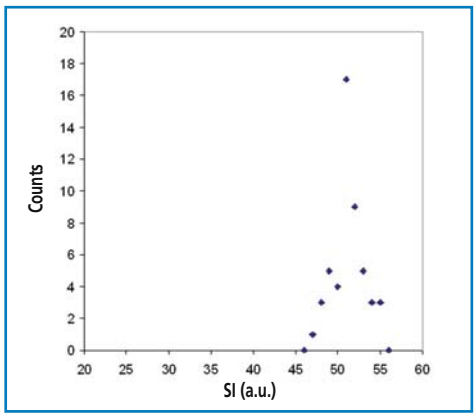


Figure 4.2. Histogram of the signal intensity (SI) measured in arbitrary units (a.u.) for the data of Table 4.3.

In order to calculate $t_{95\%}$ we now calculate the standard error of the difference. Two possibilities are open to us. We can make the hypothesis of equal variance in the two populations, or we can estimate such variance through the sample variances s^2_1 and s^2_2.

Variance pooled estimation

Homoschedasticity. If we have clear reasons to believe that the variances of the two populations are the same or we have previously demonstrated that they are not significantly different[4], we can obtain a pooled estimation of the standard deviation, s, using both the sample variances as follows:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

which represents the best estimation we have of the standard deviation of the two pooled populations. In this way, the standard error to be used for the calculation of $t_{95\%}$ is:

$$SE(m_1 - m_2) = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Substituting each value we obtain:

$$SE(50.7\text{-}39.0) = \sqrt{\frac{(50\text{-}1)1.5^2 + (50-1)1.9^2}{50+50-2}}\sqrt{\frac{1}{50}+\frac{1}{50}} = 0.57 \text{ a.u.}$$

from which:

$$t_{95\%} = \frac{(50.7 - 30.9) - 0}{0.57} = 20.4$$

From the published tables of *t* distribution with (50 – 1) + (50 – 1) = 98 degrees of freedom [ALTMAN, 1991] we obtain $p < 0.001$ ($p < 0.1\%$)[5]. Such a value has to be interpreted as follows: if the null hypothesis H_0: $m_1 = m_2$ were true, then we would have a probability less than 0.1% of observing a difference as large as the observed one (11.7 a.u.) or larger. The advent of such a low probability leads us to reject the null hypothesis and to accept the alternative hypothesis H_1. The signal intensity of the delayed enhancement using CA 2 is therefore significantly higher than that obtained using CA 1.

Heteroschedasticity. If we do not wish to make the hypothesis that the variances of the two populations are the same or we have previously demonstrated that they are significantly different from one another, the standard error of the difference is calculated as defined in Section 2.5:

[4] There is a specific statistical test, called *F* test, for verifying homoschedasticity. This test, however, is beyond the aims of this book.

[5] The reader should note that when values less than 0.001 are obtained the indication $p < 0.001$ is generally reported, which nonetheless fails to provide information on how much *p* is less than 0.001.

$$SE(m_1 - m_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where s^2_1 and s^2_2 are the two sample variances. Substituting each value we obtain:

$$SE(50.7 - 39.0) = \sqrt{\frac{1.5^2}{50} + \frac{1.9^2}{50}} = 0.34 \text{ a.u.}$$

from which:

$$t_{95\%} = \frac{(50.7 - 39.0) - 0}{0.34} = 34.5$$

From the published tables of *t* distribution with (50 – 1) + (50 – 1) = 98 degrees of freedom [ALTMAN, 1991] we obtain $p < 0.001$ ($p < 0.1\%$). This value should be interpreted in exactly the same way as in the case of homoschedasticity: if the null hypothesis H_0: $m_1 = m_2$ were true, then we would have a probability less than 0.1% of observing a difference as large as or larger than the observed one (11.7 a.u.). The advent of such a low probability leads us to reject the null hypothesis and to accept the alternative hypothesis H_1.

The reader should note that the two methods provide almost identical results and, in fact, the heteroschedasticity calculation produces even higher significance ($t_{95\%} = 34.5$ instead of $t_{95\%} = 20.4$). The greater the difference between the two variances, the larger the difference is between the results of the two calculation methods. Conversely, if the two sample variances coincide with each other, then the two methods provide exactly the same results.

Let us make another observation. We just saw that Student's *t* test may be applied both to paired and independent data and that the sole difference for the calculation of the $t_{95\%}$ coefficient is the way of obtaining the corresponding standard error of the difference $SE(m_1 - m_2)$. With independent data we also saw how to differentiate between homoschedastic and heteroschedastic data. The distinction between homoschedasticity and heteroschedasticity may also be applied when calculating the confidence interval of the difference of the two sample means. In fact, also for the calculation of the confidence interval, the sole difference between the case of paired data and independent data is the calculation of the standard error. In Section 2.5, for the sake of simplicity, we only discussed the general case of heteroschedasticity.

Analogies with confidence intervals

4.2.1. The Link with Confidence Intervals

We wanted to introduce Student's *t* test starting from the definition of confidence intervals in order to stress the strong link between the two concepts: a probabilistic conceptual bridge.

Let us consider once again the comparison between two sample means, m_1 and m_2. Suppose we compare the mean electron density in CT of a certain anatomic structure in two samples and we obtain a difference equal to 25

Hounsfield units (HU) with a 95%CI = [10, 40] HU. For the observed difference to be statistically nonsignificant, the 95%CI would need to contain zero, i.e. the expected value we would obtain if the null hypothesis were true. In the proposed example zero is not contained in the 95%CI, so we may conclude that the difference of 25 HU is statistically significant, even without performing Student's *t* test. Conversely, if the 95%CI had been, for example, [-5, 55] HU, then the observed difference should be statistically nonsignificant.

Another way of comparing two sample means is to compare the corresponding confidence intervals. In Example 4.1, the confidence intervals of the two sample means are:

$$95\%CI(50.7) = 50.7 \pm 2.010 \times 0.3 = [50.2, 52.7] \text{ a.u.}$$

$$95\%CI(39.0) = 39.0 \pm 2.010 \times 0.2 = [38.6, 41.0] \text{ a.u.}$$

The two confidence intervals do not overlap each other, so again we can conclude that the difference of 11.7 a.u. is statistically significant.

Alternatives to Student's *t* test

Therefore, we have introduced three methods of checking whether the difference between two sample means is statistically significant:

- performing Student's *t* test;
- calculating the confidence interval of the difference and verifying whether zero is contained in this interval;
- calculating the confidence intervals of the two sample means and verifying whether they are overlapping.

Although these three methods may appear different from each other, from a mathematical point of view they are all equivalent.

4.3. Comparing Three or More Sample Means: the Analysis of Variance

Three or more samples

In some circumstances one wants to compare three or more sample means, for example, in cases where the overall group of individuals is subdivided into three or more samples instead of two. Reconsidering Example 4.1, if we had introduced a third contrast agent we would have divided the initial group of 100 patients in three independent samples with sizes n_1, n_2 and n_3; the null hypothesis should be modified as follows:

$$H_0: m_1 = m_2 = m_3$$

The alternative hypothesis should involve at least one inequality between the means.

The case just described concerns the comparison of three or more independent samples. Another possibility is the comparison between three or more treatments in the same statistical sample. Let us suppose, for example, we measure renal volume by ultrasound, MR and CT and we compare the observed results

to assess any differences between the three diagnostic methods. In order to do so, a sample made up of *n* individuals could undergo all the three examinations.

Can Student's *t* test be performed for all the possible combinations? With only three sample means, for example, we could perform the *t* test to compare m_1 and m_2, m_1 and m_3, m_2 and m_3. However, this approach, although possible, is unadvisable. With both paired and independent data the right analysis method is the *analysis of variance* (ANOVA, *ANalysis Of VAriance*) but the calculation procedure is different for the two cases. In the next two sections we will see how to approach and interpret the analysis of variance, referring the reader to specialized texts for the mathematical details. Obviously, the ANOVA method may also be applied to the comparison of only two sample means: in this case it provides the same results as Student's *t* test. Lastly, note that the application of the ANOVA method also requires the verication of the conditions listed in Table 4.1.

ANOVA

4.3.1. ANOVA for Independent Groups

This type of analysis is applied to data organized as in Table 4.4.

As its name suggests, the analysis of variance consists of exploring the components of the overall observed variance. The overall variance is calculated pooling the data from all the groups, in such a way as to make a single sample whose variance is indicated by s^2 and whose mean is indicated by m. Remembering the definition of the variance, the overall variance is the sum of the squares of the differences between each sample unit and the mean, divided by the number of degrees of freedom $(n_1 + n_2 + \ldots + n_N - 1)$[6]. Now we have to introduce two other types of variability: the *within groups* variance and the *between groups* variance. The within groups variance is calculated as the sum of the squares of the differences between each statistical unit and the mean of the corresponding group.

Within groups and between groups variances

Now we shall apply the analysis of variance to the data of Example 4.1. The overall mean (calculated on all 100 patients) is m = 44.9 a.u. with an overall variance s^2 = 37.4 a.u.^{2}. From the signal intensity of each patient treated with

Table 4.4. Data organization scheme for the analysis of variance*

Group 1	Group 2	...	Group N
Individual 1	Individual 1	...	Individual 1
Individual 2	Individual 2	...	Individual 2
....	....	...	....
Individual n_1	Individual n_1	...	Individual n_1
m_1	m_2	...	m_N
s_1	s_2	...	s_N

*The measured variable has to be the same for all the groups.

[6] For the following discussion, it is convenient to express variance in this way, i.e. the sum of the squares of the differences between each sample unit and the mean, divided by the number of degrees of freedom.

CA 1 we subtract $m_1 = 39.0$ a.u., while from the signal intensity of each patient treated with CA 2 we subtract $m_2 = 50.7$ a.u.; each difference has to be squared and, last of all, these squares have to be summed. The overall sum is divided by the number of degrees of freedom (equal to $n_1 + n_2 + \ldots + n_N - N$), $100 - 2 = 98$ for Example 4.1. The between groups variance is calculated as the sum of the squares of the differences between each sample mean (m_i) and the overall mean m; this sum is then divided by the number of degrees of freedom N – 1 (i.e. the number of groups minus 1). Now we can demonstrate that the overall variance is the sum of the variances within and between groups.

The overall variance is the sum of the variances within and between groups

The logic of the analysis of variance for independent data is the following: if the null hypothesis were true, i.e. if all the sample means m_i were equal to each other, then we would think of the data in Tables 4.2 and 4.3 as all extracted from the same population and that there should be no differences between the two types of variance. In other words: *belonging to any group does not influence the overall variability*. For this reason, if the null hypothesis were true, the ratio

$$F = \frac{\text{between groups variance}}{\text{within groups variance}}$$

If the null hypothesis were true, F would tend to 1

should be equal to 1. In response to the previous statement the reader may think of the between groups variance as a measure of how much the individual means differ from the overall mean, a variability that could depend on an actual difference between the groups. In addition, the within groups variance is a measure of the variance that we would observe if all individuals belonged to the same population. Therefore, it is clear that if belonging to one group instead of another has a real effect on the corresponding mean, then F increases, and the larger the difference between the sample means, the larger the F value is.

As for Student's *t* test, the F value has to be compared with suitable published tables [Altman, 1991] from which one can obtain the corresponding *p* value, namely the probability of observing an F value as large as or higher than the observed one, if the null hypothesis were true. Since F is defined as a ratio, and since the numerator and denominator have different degrees of freedom, the F value is characterized by both the degrees of freedom, and the published tables of F values are organized in such a way as to report the most common combinations of degrees of freedom. Table 4.5 reports the results of the analysis of variance applied to Example 4.1.

The F distribution has two types of degrees of freedom

In this case, if the null hypothesis H_0: $m_1 = m_2$ were true, then the probability of observing a difference equal to or higher than $50.7 - 39.0 = 11.7$ a.u. is

Table 4.5. Result of the analysis of variance applied to Example 4.1*

Source of variation	Degrees of freedom	Sum of the squares $(a.u.)^2$	Variance $(a.u.)^2$	F	*p*
Between groups	1	3425.6	3425.6	1193.1	< 0.001
Within groups	98	281.4	2.87		
Total	99	3707.0			

*The variance is calculated as the sum of the squares divided by the number of degrees of freedom.

less than 0.1%. Since this possibility was actually observed despite the low probability, we may conclude that the null hypothesis has to be rejected and that the alternative hypothesis H_1: $m_1 \neq m_2$ may be accepted. Again, the reader should note that the *p* value is the same value obtained using Student's *t* test with the homoschedasticity hypothesis[7].

4.3.2. ANOVA for Paired Data

The ANOVA method for independent data introduced in the previous section is the natural generalization of the case of more than two sample means of Student's *t* test for independent data. Now we shall see the corresponding generalization of the *t* test for paired data.

Let us consider the following example.

Example 4.2. Comparison of four regimens of administration of contrast agent for myocardial delayed enhancement. Suppose we wish to assess the difference between the following four regimens of administration of contrast agent for delayed enhancement of the myocardium with MR imaging[8]:

- injection of a dose equal to 0.05 mmol/kg of bodyweight;
- injection of a dose equal to 0.05 mmol/kg of bodyweight followed by a second injection after ten minutes with the same dose;
- injection of a dose equal to 0.1 mmol/kg of bodyweight;
- injection of a dose equal to 0.1 mmol/kg of bodyweight followed by a second injection after ten minutes with the same dose.

For this purpose the signal intensity (in arbitrary units) is measured in a region of interest placed in the infarcted myocardium for a sample of 13 post-ischemic patients undergoing an MR examination with inversion recovery turbo-gradient-echo sequence. Data are reported in Table 4.6.

In Example 4.2 all 13 patients undergo four MR examinations, one for each regimen of administration, unlike Example 4.1, where we extracted a different sample for both contrast agents. This approach is much more powerful than that used for independent data, because it allows us to focus on the differences within each individual of the sample, differences due to the variable placed in columns[9]. For the sake of clarity, we will not verify the conditions for the application of the ANOVA analysis.

Between and within subjects variances

The reader may easily see that the distinction we introduced in the previous section regarding between groups variance and within groups variance no longer holds

[7] To be rigorous, when the F numerator has only one degree of freedom (i.e. when we are comparing only two sample means) it is $F = t^2$.

[8] Note that a study like this has been really performed by our research group. However, for ethical reasons, to avoid the need of four examinations and four contrast injections in the same patient, each patient underwent only two exams, each of them with two sequential contrast administrations (0.05 mmol/kg followed by 0.05 mmol/kg and 0.1 mmol/kg followed by 0.1 mmol/kg, in randomized order of priority).

[9] Often, the variable placed in a column is called *factor*.

Table 4.6. Signal intensity for the four regimens of administration of contrast agent of Example 4.2*

Patient	0.05 mmol/kg	0.05+0.05 mmol/kg	0.1 mmol/kg	0.1+0.1 mmol/kg	m	s
1	51.0	48.5	32.1	45.1	44.2	8.4
2	27.5	57.2	55.5	75.2	53.9	19.7
3	66.9	45.7	54.0	81.6	62.1	15.7
4	15.2	54.6	39.4	49.8	39.8	17.6
5	48.4	49.1	43.7	52.1	48.3	3.5
6	12.1	24.3	45.2	49.9	32.9	17.8
7	29.1	30.6	43.3	75.3	44.6	21.5
8	38.6	34.0	25.2	50.3	37.0	10.5
9	51.6	36.2	37.1	26.2	37.8	10.5
10	11.6	37.0	22.7	36.3	26.9	12.1
11	41.6	26.9	30.6	28.1	31.8	6.7
12	38.2	42.1	41.0	38.7	40.0	1.9
13	24.3	52.8	29.0	53.6	39.9	15.5
m	35.1	41.5	38.4	50.9	41.5	
s	17.0	10.8	10.2	17.5	9.4	

*Data are signal intensities expressed in arbitrary units.
Note that the occurence of different standard deviation (s) should be considered as a contraindication to the use of parametric ANOVA. The reader can retain this table only as an example to show the logic of the method.

and that a new distinction has to be made regarding *between subjects* variance and *within subjects* variance. This difference depends on the data symmetry (see Table 4.6) which allows for the calculation of the mean and of the variance both *in a horizontal* and *in a vertical* way. However, while the overall variance with independent data is the sum of the between groups and within groups variances, with paired data, in addition to the between subjects and within subjects variances there is also a *residual variance*. Moreover, for Example 4.2 the within subjects variance may also be considered as a kind of *between regimens of administration variance*. The question is: "Does the mean signal intensity depend on the regimen of administration of the contrast agent?" In other words: "Are the differences between the means calculated for each regimen of administration statistically significant?"

For the mathematical details the reader should refer to specialized texts. Here we report the results of the ANOVA method for the data in Example 4.2, as provided by a common statistical software package (Table 4.7).

Now we shall see how to interpret the data in Table 4.7.

As usual, each single variance is calculated by dividing the corresponding sum of squares by the degrees of freedom, while the F value is obtained by dividing the corresponding variance by the residual variance. From the published tables [ALTMAN, 1991] of the F distribution with 12 and 36 degrees of freedom and with 3 and 36 degrees of freedom we obtain the p value. The first p value ($p = 0.027$) indicates that the differences between the patients (in terms of signal intensity) are statistically significant; this result is of little interest and does not answer the question we posed. The second and more important p value ($p = 0.016$) indicates that also the differences between the four regimens of administration of the contrast agent are statistically significant, i.e. the mean signal intensity depends on dose and administration regimen of the contrast agent.

Table 4.7. Results of the analysis of variance for Example 4.2*

Source of variation	Degrees of freedom	Sum of the squares (a.u.)^2	Variance (a.u.)^2	F	*p*
Subjects	12	4245.3	353.8	2.30	0.027
Regimens of administration	3	1820.5	606.8	3.94	0.016
Residuals	36	5540.9	153.9		
Total	51	11606.7			

*The variance is the ratio between the sum of the squares and the number of degrees of freedom. The F value is the ratio between the corresponding variance and the residual variance. a.u. = arbitrary unit.

4.4. Parametric Statistics in Radiology

Parametric techniques all have the same mathematical logic

Due to the underlying assumptions related to normal distribution, parametric techniques have a general meaning in biostatistics. In fact, it can be demonstrated that all parametric techniques belong to the same mathematical scheme. Moreover, they introduce the general conceptual scheme of the hypothesis tests. These are very powerful tests, able to demonstrate the significance of small differences and/or of differences found in small samples.

However, this power depends on stringent conditions, which include:

– the variable type (necessarily continuous);
– the data distribution (necessarily normal);
– the variance (which when comparing two or more sample means are necessarily not statistically different from each other).

Therefore, parametric techniques *depend on distribution*.

In radiologic research we commonly measure categorical or ordinal variables. Moreover, even when measuring continuous variables, we often have non-normal distributions (mean and median are far from each other) or the sample size does not allow us to demonstrate if we are dealing with normal distributions. Therefore, rarely may we correctly use parametric methods; even rarer are the cases in which the conditions for their application are verified (this verification should always be performed by statisticians).

Despite their lower power than the corresponding parametric tests, non-parametric statistical tests are more frequently used in scientific radiologic studies because:

– they are able to handle non-continuous variables;
– they allow the radiologist to not verify the conditions described above.

However, the fundamental concepts of parametric statistics need to be understood in order to understand those of non-parametric statistics.

References

Altman DG (1991) Practical statistics for medical research. London: Chapman & Hall

American College of Radiology (2003) ACR breast imaging reporting and data system (BIRADS): Breast imaging atlas, 4th ed. Reston, VA

Soliani L (2007). Statistica applicata alla ricerca e alle professioni scientifiche. Manuale di statistica univariata e bivariata. Parma: Uninova-Gruppo Pegaso; 6:2-8 (http://www.dsa.unipr.it/soliani/soliani.html)

5

Non-Parametric Statistics

Mathematicians are like Frenchmen:
whatever you say to them
they translate into their own language
and forthwith it is something entirely different

JOHANN WOLFANG GOETHE

The birth of *non-parametric statistics* is historically related to the solution of methodologic problems in experimental psychology. It was Stanley S. Stevens (1906-1973) who solved the question about the inappropriate use of measurement scales; he also proposed a new classification that gave rise to the distinction between nominal scales, rank scales, interval scales and continuous scales, a distinction we introduced in Chapter 2 (see Table 2.1). Based on this, *behavioral science statistics* was developed in the 1940s, in part thanks to other researchers such as Quinn McNemar (1900-1986), Frederick Mosteller (b., 1916) and Anthony W.F. Edwards (b., 1935), with a large use of *non-parametric methods* [CARACCIOLO, 1992]. Moreover, non-parametric statistics is also the result of a broader discussion between the founding fathers of *Theoretical Statistics* and the founding fathers of *Modern Statistics* (see Introduction to Chapter 4).

Since the studies of Francis Galton (1822-1911), statisticians have extensively applied the hallmarks of Gaussian distribution. In practice, they performed calculations and arrived at conclusions without verifying the necessary conditions for the use of parametric methods. They made many relevant errors such as using the analysis of variance with dichotomous variables.

Non-parametric statistics poses fewer bonds to data distribution

The definition of non-parametric statistics is based on the absence of bonds related to normal distribution parameters. The logical link is the following: parametric statistics is based on Gaussian distribution features which, in turn, depends on only two parameters, the mean and standard deviation. Since the new methods do not impose any conditions on the distribution shape, they are called *non-parametric tests*, because they are not based on the mean and standard deviation.

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

This definition may be confusing for the reader, especially if we consider that, when going into mathematical details, non-parametric statistics also uses many indices and parameters. A more correct terminological framework sees the absence of *a priori* assumptions regarding distribution shape. For example, a more appropriate term may be *distribution-free*. However, the term "distribution free" suggests the data distribution type has no importance. Actually, many non-parametric methods also require that some less stringent assumptions on distribution shape be satisfied. At any rate, regardless of the above considerations, *these tests are now most commonly referred to as "non-parametric tests"*.

Distribution-free

Advantages of non-parametric stastics

One very important advantage of these tests is their versatility – in fact they have a wide range of applications. As stated in Chapter 4, we may only use parametric statistics with continuous data. This limitation is due to the type of the mathematical calculations that the data undergo, starting from the calculation of the mean and standard deviation; moreover, this limitation reduces the number of parametric tests. On the other hand, non-parametric statistical tests can be used to analyze all types of variables and measurement scales and this important feature enabled the development of many statistical tests, each for a specific task. This latter characteristic has an important impact on radiologic research, where all types of variables appear. Another valuable advantage of non-parametric statistics is its power with small samples.

This chapter, unlike the previous one, does not describe the mathematical details of these tests. This decision was made on the one hand to give more space to conceptual aspects, and on the other to provide the reader with a kind of *practical guide*, i.e. a reference book to establish which is the more suitable test on a case-by-case basis. Although for each test we briefly describe the calculation procedure, our advice is to use dedicated statistical software. The discussion scheme is organized based on the various circumstances one can come across in practice. For each test one or more examples are presented. We refer to the systematic classification proposed by Sidney Siegel and N. John Castellan Jr [Siegel and Castellan, 1992].

5.1. One Sample with Two Paired Measurements

The comparison between *pairs of dependent measurements (paired data), typically two observations within the same individuals*, may be performed by many non-parametric tests. Examples of this include patients undergoing two different imaging modalities or with two different techniques of the same imaging modality (high and low spatial resolution, with and without contrast agent, two different MR sequences, etc.) or before and after a certain therapy with the same modality and technique.

5.1.1. Variables Measured with Dichotomous Scale

McNemar test

In this case the right test is the *McNemar test on changes* or, with small samples, the *binomial test*. These tests can be applied every time the measurement has a dichotomous trait such as *yes/no* or *positive/negative*. This is the typical case of studies of diagnostic performance, because a radiologic examination

provides its result in terms of the presence (positive result) or absence (negative result) of disease. It is also possible to use the McNemar test with variables measured with higher level measurement scales, after being dichotomized distinguishing values lower and higher than a certain threshold. In the following example we consider the simple case of the presence/absence of a certain finding in two mammograms performed by the same patient but with two different breast compression techniques.

Example 5.1. One hundred women who undergo a periodically scheduled mammography are enrolled in a prospective study aimed at evaluating a new breast compression system called *biphasic compression* (BC). With this technique the compression plate initially comes down at an angle of 22.5° to the film cassette and then finishes parallel to it. Following a randomization protocol, 25 women undergo the craniocaudal (CC) view of the right breast twice, once with the standard monophasic compression (MC) system and once with the biphasic compression; similarly, 25 women undergo the CC view of the left breast twice; 25 women undergo the mediolateral-oblique (MLO) view of the right breast twice; 25 women undergo the MLO view of the left breast twice. Moreover, the performing order of the two compression techniques and the execution of the mammogram pairs by Radiographer 1 and Radiographer 2 are also randomized. During the examinations measurements are made of the compressed breast thickness and the distance between the anterior nipple surface and the posterior margin of the film for the CC view and the distance between the anterior nipple surface and the anterior margin of the pectoral muscle for the MLO view (*posterior nipple line*). The visibility of the pectoral muscle for the CC view and that of the submammary fold for the MLO view serve as quality index [Sardanelli et al, 2000].

The reader will have noted that example 5.1 deals with different variables. The breast thickness and the exposure parameters are continuous variables; the visibility of the pectoral muscle and the submammary fold are dichotomous variables.

A part of the results is summarized in Table 5.1.

Now let us here consider the dichotomous variable. As shown in Table 5.1, for the CC view the pectoral muscle was visible in 27 out of 50 mammograms (54%) performed with biphasic compression and in 17 out of 50 mammograms (34%) performed with standard compression. For the MLO view, the submammary fold was visible in 45 out of 50 mammograms (90%) and in 36 out of 50 mammograms (72%), respectively. The McNemar test demonstrates a significant difference in favor of the biphasic compression both for the CC view ($p = 0.006$) and for the MLO view ($p = 0.022$).

The McNemar test only considers changes

Procedure. The McNemar test only considers changes, i.e. the sample units whose two measures are different from each other. The null hypothesis expects that the number of changes should be equiprobable in both directions and, therefore, that half of the discrepancies concern individuals that pass from "positive" to "negative" and that half pass from "negative" to "positive" (expected discrepancies). The concordances, i.e. the statistical units whose

Table 5.1. Results of the study of the Example 5.1 (part one)

Findings With Breast BC versus Standard MC at X-ray Mammography

Findings	BC	MC
Posterior nipple line distance (cm)		
CC*		
Mean ± SD	10.5 ± 2.3	10.2 ± 2.2
Range	6.0 – 15.3	6.0 – 14.6
MLO†		
Mean ± SD	11.0 ± 2.1	10.8 ± 2.1
Range	6.4 – 15.1	6.2 – 15.1
Pectoral muscle (n = 50)‡	27 (54)	17 (34)
Inframammary fold (n = 50)§	45 (90)	36 (72)
Thickness of compressed breast (cm)		
CC‖		
Mean ± SD	4.8 ± 1.1	4.7 ± 1.1
Range	2.2 – 7.2	1.9 – 7.0
MLO#		
Mean ± SD	5.1 ± 1.1	4.8 ± 1.1
Range	2.0 – 8.2	2.0 – 8.0

* Difference, 0.35 ± 0.04 (mean ± standard error); $P < .001$ (Wilcoxon matched pairs signed rank test).
† Difference, 0.34 ± 0.05; $P = .002$ (Wilcoxon).
‡ Data are the number of such findings. Numbers in parentheses are percentages. $P = .006$ (McNemar test).
§ $P = .022$ (McNemar).
‖ Difference, 0.20 ± 0.04; difference not significant (Wilcoxon).
Difference, 0.22 ± 0.05; difference not significant (Wilcoxon).

BC = biphasic compression; MC = monophasic compression. From: Sardanelli F, Zandrino F, Imperiale A et al (2000) Breast biphasic compression versus standard monophasic compression in x-ray mammography. Radiology 217:576-580 (with permission of the copyright owner [RSNA]).

judgment does not change, do not enter into the calculation. If the observed discrepancies differ from the expected ones other than due to chance, then the test is significant.

In the example, there were 22 cases for the CC view in which the pectoral muscle was not visible with both techniques and 16 cases in which it was visible; in 11 cases the pectoral muscle was visible with biphasic compression and not visible with standard compression and, lastly, in 1 case it was visible with standard compression and not visible with biphasic compression. Mathematically, the test only considers the 12 discrepancies, 11 in favor of the biphasic compression and 1 in favor of the standard compression, and provides $p = 0.006$.

Likewise, for the MLO view there were 3 cases in which the submammary fold was not visible with both techniques and 34 cases in which it was visible; in 11 cases the submammary fold was visible with biphasic compression and not visible with standard compression; and, lastly, in 2 cases it was visible with standard compression and not visible with biphasic compression. Mathematically, the test only considers the 13 discrepancies, 11 in favor of biphasic compression and 2 in favor of standard compression, and provides $p = 0.022$.

Since the number of the discrepancies was very small, the p value is calculated using binomial distribution (*binomial test*).

Example 5.2. The application of the McNemar test to the general case is illustrated in the following theoretical example.

The diagnostic modalities A and B are compared with the reference standard for a sample of 200 cases (patients or lesions) for the detection of disease D. At the reference standard, 100 are positive cases and 100 are negative cases. Diagnostic modality A has a sensitivity equal to 78% (78/100), while B has a sensitivity equal to 58% (58/100); A has a specificity equal to 68% (68/100), B equal to 85% (85/100). In order to establish if A is actually more sensitive than B and if B is actually more specific than A, we have to evaluate concordances and discrepancies, case by case.

Let us consider the 100 positive cases at the reference standard:

– in 45 cases both A and B are *true positives*;
– in 8 cases both A and B are *false negatives*;
– in 34 cases A is *true positive* and B is *false negative*;
– in 13 cases A is *false negative* and B is *true positive*.

The McNemar test only considers the 47 discrepancies and provides $p = 0.004$: diagnostic modality A is significantly more sensitive than B for disease D.

Let us consider the 100 negative cases at the reference standard:

– in 7 cases both A and B are *false positives*;
– in 60 cases both A and B are *true negatives*;
– in 25 cases A is *false positive* and B is *true negative*;
– in 8 cases A is *true negative* and B is *false positive*.

The McNemar test only considers the 33 discrepancies and provides $p = 0.005$: diagnostic modality B is significantly more specific than A for disease D.

In both cases, the sample size was large enough to perform the McNemar test instead of the binomial test.

The diagnostic accuracy of modality A is 73.5% (147/200) and the diagnostic accuracy of modality B is 71.5% (143/200). If we want to establish if A is more accurate than B we have to consider the number of cases in which A and B agree with both the reference standard and each other:

– in 105 cases (45 + 60) A and B are either *true positives* or *negatives*;
– in 15 cases (8 + 7) A and B are either *false positives* or *negatives*;
– in 42 cases (34 + 8) A is either *true positive* or *negative* and B is either *false positive* or *negative*;
– in 38 cases (13 + 25) A is either *false positive* or *negative* and B is either *true positive* or *negative*.

The McNemar test only considers the 80 (42 + 38) discrepancies and provides $p = 0.738$: we do not have evidence to reject the null hypothesis which states "A is as accurate as B for disease D". In other words, the accuracy difference between the diagnostic modalities A and B is not significant.

If the number of expected discrepancies is lower than 5, use the binomial test

Comment. As often happens, the approximations made when performing a statistical test no longer hold with small sample sizes. To use the McNemar test the number of expected discrepancies should be at least equal to 5. Otherwise, it is more correct to use the binomial test. The reader should note that both the

McNemar test and the binomial test apply to the same data type and, therefore, *many statistical software packages automatically choose which test to perform based on the sample size*. For this data type there is no corresponding parametric test and therefore the *power* of the McNemar test cannot be estimated. For a broader understanding of the logical-mathematical setting of this test, the reader is invited to consult McNemar [McNemar, 1969].

An important point which needs to be made regards the use of the McNemar test for comparing diagnostic performance indices. Using this test for comparing paired data is quite simple. In its original version, the test cannot be applied to compare predictive values, since the number of the positive test findings is included in the denominator of the positive predictive value while the number of the negative test findings is included in the denominator of the negative predictive value. For this reason, the denominator may be different for the two diagnostic modalities and the comparison between them cannot be based on a complete series of paired data. A modified McNemar test for the comparison of the predictive values involves mathematically complex procedures which go beyond the aims of this book. The interested reader may consult Leisenring et al. [Leisenring et al, 1997; Leisenring and Pepe, 1998].

5.1.2. Variables Measured with Ordinal Scales

Sign test

In this instance the test to be performed is the *sign test*, based on the direction (positive or negative) of the changes of the pairs. The sign of the change may be assessed thanks to the nature of the ordinal variable.

> **Example 5.3.** To assess the image quality of the aortic valve at multidetector CT with retrospective ECG gating with and without iodinated contrast agent, 25 patients are studied prior to surgery. Two radiologists evaluate image quality by consensus using an ordinal scale with the following scores: 1 = nondiagnostic quality; 2 = poor but diagnostic quality; 3 = good quality; 4 = excellent quality. They then evaluate the definition of the aortic valve morphology using an ordinal scale with the following scores: 1 = possibly correct definition; 2 = probably correct definition; 3 = definitely correct definition. The authors of this study report the details of the criteria used for assigning both the scores. The sign test provides a highly significant difference towards CT with iodinated contrast agent rather than CT without contrast agent both for image quality ($p = 0.004$) and for the definition of the aortic valve morphology ($p = 0.006$) [Willmann et al, 2002].

Procedure. For each data pair we need to established which value is higher than the other. The pairs with identical values are not included in the calculation. With regard to Example 5.3 for the definition of aortic valve morphology, we may hypothesize the data distribution shown in Table 5.2.

We may hypothesize 11 pairs in favor of CT with iodinated contrast agent, 1 pair in favor of CT without contrast agent, and 3 pairs with the same score. The sign test only considers the 12 pairs with different scores in the two techniques and provides $p = 0.006$.

Table 5.2. Definition of aortic valve morphology for Example 5.3 [Willmann et al, 2002]*

	Morphology definition		
	1	2	3
CT without iodinated contrast agent	9	3	3
CT with iodinated contrast agent	0	5	10

* The authors report the distribution of 15 patients who underwent CT without contrast agent and 25 patients who underwent CT with contrast agent. For the sake of simplicity the distribution of only the 15 patients who underwent both techniques is hypothesized here.

Example 5.4. To assess the impact of ECG gating on the image quality of thin-slice CT of the lung, 45 patients prospectively undergo the examination, with and without ECG gating. Three radiologists evaluate by consensus the image quality of the superior lobes, the central or lingula lobe and the two inferior lobes using a five point ordinal scale from 1 (worst) to 5 (best) for the presence of noise, motion artifacts and overall diagnostic assessment. The sign test with *Bonferroni's correction*[1] demonstrates no significant differences for the presence of noise for any one of the lobes; it demonstrates a significant difference for the presence of movement artifacts for the central lobe, the lingula lobe and both the left and right inferior lobes ($p < 0.004$); it demonstrates a significant difference for the overall diagnostic assessment only for the inferior left lobe ($p < 0.004$) [Bohem et al, 2003].

Bonferroni's correction

Comment. Data couples may also come from two different individuals, belonging to different populations. However, it is important for the data to be matched in a homogenous way, thereby canceling the influence of other factors. Also in this case the calculation is made only with individuals with two different values; pairs of identical values are not included in the calculation. If we use the sign test on data for which the Student *t* test is applicable, its power is 95% with N = 6, where N is the number of couples with different values; this power decreases as N increase, up to 63% (asymptotic power).

Sign test power is between 63% and 95%

When dealing with interval or rational variables and with small N values using the *test of the permutations* is advisable. This test takes into account all the possible observed differences (equal to 2^N) through combinatory calculation. It has 100% power. There is also a modification of this test for independent data.

Test of the permutations

[1] Bonferroni's correction is a very conservative method which is applied when making multiple paired comparisons (in this case 18 comparisons: 3 comparisons for each of the 6 lung lobes). It consists of multiplying each of the observed *p* values by the number of comparisons. The correction for multiple comparisons is beyond the aims of this book, but it is important the reader is aware of it. The problem arises from the need to take into account that choosing the α error as 0.05 produces a 5% possibility of obtaining a false positive. For Example 5.4, with 18 comparisons, we should have a very high probability of obtaining a false positive without Bonferroni's correction. However, some authors consider Bonferroni's correction is a too conservative correction. For further details refer to Douglas G. Altman [Altman, 1991].

5.1.3. Variables Measured with Interval or Rational Scales

Wilcoxon test

If the variable is measured with an at least interval measurement scale and if its distribution may be considered as continuous, then the *Wilcoxon signed rank test*, commonly called *Wilcoxon test* may be used.[2] This test compares the medians of the two samples. Unlike the sign test, in the Wilcoxon test the larger the difference between the two values is, the larger the weight for the calculation of that difference.

Example 5.5. For the sake of simplicity let us refer to Example 5.1 and Table 5.3 regarding the length of the *posterior nipple line*. With the craniocaudal (CC) view the difference between biphasic and standard compression is 0.35 ± 0.04 cm (mean ± standard error); 0.34 ± 0.05 for the medio-lateral-oblique (MLO) view. Using the Wilcoxon test we get $p < 0.001$ for the CC view and $p = 0.002$ for the MLO view. This means that biphasic compression significantly increases the imaged breast amount. *Note how a difference as small as few millimeters may not only be statistically significant, but also clinically relevant*. In fact, such a small difference for a one-dimensional measurement implies a larger difference for the two-dimensional measurement of breast surface and even more for the three-dimensional measurement of breast volume. This allows for a more extended analysis of the region placed between the gland and the pectoral muscle, a typical site of breast carcinoma. On the other hand, as we may see in Table 5.2, there are no statistically significant differences between the two compression systems in terms of breast thickness, neither for the CC nor the MLO views.

Statistically significant and clinically relevant difference

Procedure. The Wilcoxon test takes into account the absolute difference between the two measurements of each statistical unit. Such differences are associated with ranks and each rank obtains a sign (positive or negative) based on the initial difference sign. If one or more statistical units have a zero difference, then those units are excluded from the calculation and the sample size decreases. The null hypothesis states that the observed difference is not significant and that the two medians coincide with each other. Thus, separately adding the positive and the negative ranks we would have two identical sums. The test is significant if the difference between the above sums is larger than that which can be explained with normal statistical fluctuations.

The null hypothesis states that the two distribution medians coincide with each other

Comment. Because it is non-parametric in nature, the Wilcoxon test does not require normal data distribution. Nevertheless, it does need symmetric distribution for the differences. If this latter requirement is not verified, the data may be transformed by generating a new, more symmetric distribution.

The power of the Wilcoxon test is 95.5%

If used on data that verify the conditions for applying the Student *t* test, the power of the Wilcoxon test is equal to 95.5%. For more details the reader may consult Conover [Conover, 1999].

[2] The reader should pay attention to the terminology. The name Wilcoxon is associated with three different statistical tests, together with the names of Mann and Whitney. The test reported here is the most common.

5.2. Two Independent Samples

5.2.1. Variables Measured with Nominal or Ordinal Scales

Two independent samples, *even of different sizes*, may be produced by the random extraction from two populations or by the random association of two *treatments*. This happens, for example, when comparing the performance of two diagnostic techniques for a given disease, one performed in one patient group and the other in a different patient group.

χ^2 test

With categorical (nominal or ordinal) data the right test to be performed is the *chi-square (χ^2) test*. This test compares all the features of the distributions from which the two samples are extracted (central tendency, spread, symmetry, etc.). *It is the general test to be performed when comparing occurrence frequency in different groups.* If we are dealing with a dichotomous variable and with small sample sizes, the *Fisher exact test* may be used. The starting point is a typical 2×2 contingency table. Through the technique of the combinatory calculation, the Fisher exact test arrives at the exact probability of obtaining the observed frequency distribution. If the total number of observations (N) is larger than 20, the calculation may become prohibitive. In these cases the χ^2 test may be used.

Example 5.6. Let us reconsider Example 5.1 for assessing the performance of the two radiographers who performed the mammographies, as reported in Table 5.3. The χ^2 test or the Fisher Exact test is used to assess for any differ-

Table 5.3. Results of the study of Example 5.1 (part two)

Performance With BC versus Standard MC at X-ray Mammography

Performance	BC	MC
Posterior nipple line distance (cm)		
CC*		
Radiographer 1		
Mean ± SD	10.7 ± 2.3	9.9 ± 2.2
Range	6.0 – 15.0	6.0 – 14.3
Radiographer 2		
Mean ± SD	10.3 ± 2.3	10.4 ± 2.3
Range	6.1 – 15.3	6.3 – 14.6
MLO†		
Radiographer 1		
Mean ± SD	11.0 ± 2.0	10.8 ± 2.2
Range	6.6 – 14.0	6.2 – 14.0
Radiographer 2		
Mean ± SD	10.9 ± 2.3	10.8 ± 2.2
Range	6.4 – 15.0	6.5 – 15.1
Pectoral muscle (n = 25)‡		
Radiographer 1	12 (48)	11 (44)
Radiographer 2	15 (60)	6 (24)
Inframammary fold (n = 25)§		
Radiographer 1	22 (88)	15 (60)
Radiographer 2	23 (92)	21 (84)

* BC, P = .449 (not significant [NS]) (Mann-Whitney U test); MC, P = .398 (NS).
† BC, P = .899 (NS) (Mann-Whitney); MC, P = .712 (NS).
‡ BC, P = .395 (NS) (χ^2 test); MC, P = .135 (NS).
§ BC, P = .99 (NS) (Fisher exact test); MC, P = .059 (NS) (χ^2 test).

BC = biphasic compression; MC = monophasic compression. From: Sardanelli F, Zandrino F, Imperiale A et al (2000) Breast biphasic compression versus standard monophasic compression in x-ray mammography. Radiology 217:576-580 (with permission of the copyright owner [RSNA]).

ence in performance of the two radiographers. All differences in visualizing the pectoral muscle in the CC view and the submammary fold in the MLO view were not statistically significant, even if Radiographer 1 showed better improvement using the MLO view while Radiographer 2 showed better improvement using tho CC view.

Procedure. The χ^2 test can be applied to data organized in contingency tables. In Table 5.3 the two compression systems (BC and MC) are placed into columns, while in the rows we find the two radiographers[3]; the four cells of this contingency table report the number of times in which the radiologic sign (the pectoral muscle and the submammary fold) was detected (with percentages in parentheses). The χ^2 test compares the frequency of each cell with the corresponding expected frequency, with the latter calculated by hypothesizing that there is no relationships between the two variables. The further the observed frequencies are from the expected ones, the more significant the test is.

The Fisher exact test with small samples

Comment. The χ^2 test can be applied without verification of any assumptions for samples whose sizes are not too small ($N > 40$). With small samples the expected frequencies may be too small. In this case, the Fisher exact test should be used, since it differs from the χ^2 test in the way the expected frequencies are calculated. The Fisher test is called "exact" because it uses the exact formula in calculating the expected frequencies, instead of the approximated formula as in the χ^2 test. The choice between the two tests is based on the following criteria [Siegel and Castellan Jr, 1992]:

- for $N \leq 20$ always use the Fisher exact test;
- for $20 < N < 40$ the χ^2 test may be used if all the expected frequencies are larger than 5; if one or more of the expected frequencies is smaller than 5 the Fisher exact test may be used;
- for $N \geq 40$ always use the χ^2 test.

Yate's correction for continuity

The reader should note that both the χ^2 and the Fisher test apply to the same data type and that many statistical software packages automatically choose the right test to be performed based on the sample size; if the χ^2 test is chosen, these computer programs also apply the *correction for continuity* introduced by Yates in 1934 [Yates, 1934]. Moreover, unlike the Fisher exact test, the χ^2 test may also be applied to contingency tables with more than 2 rows or more than 2 columns [Armitage and Berry, 1994].

The χ^2 test has no corresponding parametric test; so there is no sense speaking about its power.

5.2.2. Variables Measured with Interval or Rational Scales

Mann-Whitney *U* test

In these cases the *Mann-Whitney U test* may be used, which compares the medians of the two samples.

[3] Note that the opposite choice, with the Radiographers in the columns and the compression systems in the rows, is equivalent.

Example 5.7. Let us reconsider Table 5.3 of Example 5.1 in order to evaluate the difference between the performance of the two radiographers with regard to the posterior nipple line. With both the CC and the MLO views, the Mann-Whitney *U* test demonstrates no significant differences between the two radiographers.

Procedure. The calculation procedure consists of combining the two groups of data (X, Y) into a single sample and associating ranks. For each X value one has to count how many values of the single sample are lower than X: $U(YX_i)$; then one has to calculate the mean of $U(YX_i)$ based on all the X values. The same calculation is made for Y, obtaining the mean of $U(XY_i)$ for all the Y values. Thus we have two variability indices, one for $U(YX_i)$ and the other for $U(XY_i)$. Combining these indices we calculate *U* from which we obtain the *p* value.

Comment. The Mann-Whitney *U* test, together with the Wilcoxon test for paired data, verifies the null hypothesis that the two medians coincide with each other, without hypotheses regarding the variances and distribution type. Compared to the Student *t* test, the power of the Mann-Whitney *U* test is about 95% even for small samples. The comparison between two independent samples whose corresponding populations have different variances is known as *Behrens-Fisher problem* and the non-parametric statistical tests for this circumstance have only recently been introduced [Conover, 1999].

The power of the Mann-Whitney *U* test is close to 95%

5.3. Three or More (k) Dependent Samples[4]

5.3.1. Variable Measured with Dichotomous Scale

In this case the test to be performed is the *Cochran Q test*.

Cochran *Q* test

Example 5.8. Let us suppose we have to compare the performance of k = 4 radiology residents (each one attending a different year of the study course) for the detection of a certain radiologic sign in a sample of N individuals all undergoing the same examination. The goal is not the evaluation of the relative sensitivity and specificity (requiring a *standard of reference* which, in this case, could be a senior radiologist) but rather the assessment of possible perception differences among the four residents.

Procedure. The data have to be placed in a table with N rows (in which we will put the sample individuals) and with four columns (in which we will put the reports of the four residents), in exactly the same way the ANOVA method for paired data is done. The Cochran *Q* test verifies if the outcomes of the four residents significantly differ from each other. The procedure calculates a coefficient, denoted *Q*, by hypothesizing that there are no differ-

[4] The statistical tests here reported may also be used when k = 2.

ences among the four residents. This coefficient has an approximated χ^2 distribution with $k - 1 = 4 - 1 = 3$ degrees of freedom. The test is significant if Q is larger than a certain critical value.

The Cochran Q test is an extension of the McNemar test

Comment. The Cochran Q test is the extension of the McNemar test for more than two dependent samples. It cannot be used with a small sample size. As a rule, the sample size should be:

– $N \geq 4$;
– $N \cdot k \geq 24$.

Since there is no corresponding parametric test, the power of the Cochran Q test cannot be established.

5.3.2. Variables Measured with Ordinal, Interval or Rational Scale

Friedman test

In this instance the test to be performed is the *Friedman test* or *two-way rank ANOVA*.

> **Example 5.9.** Let us suppose that N patients with myocardial infarct undergo contrast enhanced cardiac-MR to measure the area showing myocardial delayed enhancement. This measurement is repeated four times every five minutes: the goal is to check for any significant difference.

Procedure. The data have to be arranged as for the ANOVA method. The Friedman test verifies if all four medians coincide with each other, against the alternative hypothesis which states at least one inequality. Each value in the rows is converted into ranks from 1 to 4: if the null hypothesis is true, all of the ranks will have the same frequency in the four columns and the mean ranks will coincide with each other. The calculation procedure obtains a coefficient whose distribution is known. The test is significant if this coefficient is larger than a certain critical value.

The Friedman test has a power between 64% and 91%

Comment. When $k \geq 5$ or if the sample size is very large, the coefficient calculated using the Friedman test has an approximated χ^2 distribution with $k - 1$ degrees of freedom. Compared to the analysis of variance, the Friedman test has a power equal to 64% for $k = 2$ and it increases as k increases, until the asymptotic value of 91%. For more details the reader may consult Conover [Conover, 1999].

5.4. Three or More (k) Independent Samples[5]

5.4.1. Variables Measured with Nominal or Ordinal Scale

In these circumstances the test to be performed is the χ^2 *test*.

[5] The statistical tests reported here may also be used when $k = 2$.

Example 5.10. Let us suppose that N patients with carotid artery stenosis undergo contrast-enhanced MR angiography and post-contrast scans to assess carotid plaque enhancement. We divide the sample into two groups: a group consisting of all the patients showing plaque enhancement and the other group, consisting of the patients who do not show plaque enhancement. In this way we divide the whole sample into two subgroups which are not necessarily equal in size. In addition, we evaluate the degree of stenosis with the following score:

– 0, stenosis less than 30%;
– 1, stenosis between 30% and 75%;
– 2, stenosis greater than 75%.

We would like to know whether there is a possible relationship between the two variables: plaque enhancement and degree of stenosis[6].

The generalized χ^2 test

Procedure. In this case we simply apply an extension of the χ^2 test already used for two independent samples. However, the data now has to be structured in a 2×3 contingency table, because the degree of stenosis may take three different values on an ordinal scale. Again, the χ^2 test compares the observed frequencies of each cell of the table with the corresponding expected frequencies, with these latter being calculated by hypothesizing that there is no relationship between the two variables. The larger the difference between the observed and the expected frequencies, the more significant the χ^2 test is.

Comment. The same comments on the sample size and on the power we saw for the classic χ^2 test are valid. Nevertheless, the Fisher exact test is not applicable for contingency tables other than the classic 2×2 type. If the expected frequency is less than 5 in over 20% of the total cell number, it is advisable to combine the cells in order to reduce its total number.

5.4.2. Variables Measured with Interval or Rational Scale

Kruskal-Wallis

In this instance the test to be performed is the *Kruskal-Wallis test* or *one-way rank ANOVA*.

Example 5.11. Let us reconsider Example 5.6 with a modification of the number of radiographers: now we will compare the performance of three instead of two radiographers in terms of the posterior nipple line.

Procedure. The data have to be converted into a single series of ranks. For each radiographer, the sum of the ranks and their mean have to be calculated. This test uses a coefficient, denoted *KW*, whose distribution is known. The test is significant if this coefficient is larger than a certain critical value.

[6] The reader should note that this approach is equivalent to testing for any differences in terms of the degree of stenosis between the two groups.

The Kruskal-Wallis test has a power close to 95%

Comment. When $k > 3$ and when the number of individuals is larger than 5 for each group, the coefficient calculated using the Kruskal-Wallis test has an approximated χ^2 distribution with $k - 1$ degrees of freedom. The power of this test tends to 95.5% when compared with the ANOVA method. For more details the reader may consult Conover [Conover, 1999].

5.5. Some Considerations Regarding Non-Parametric Tests

In this chapter we introduced the non-parametric statistical tests most commonly used for radiologic research. For schematization purposes we briefly indicated the calculation procedure, without weighing down the discussion with mathematical details which can be easily found in specialized texts.

Now we will provide some specific considerations for some of these tests.

We stated that in most cases the null hypothesis (H_0) takes into account the medians of the two or more samples that are compared to each other. However,

Table 5.4. Commonly used parametric and non-parametric statistical tests

Goal	Variables and measurement scales	Non-parametric methods	Parametric methods
To compare two dependent samples	Categorical with dichotomous scale	McNemar test Binomial test	NA
	Categorical with ordinal scale	Sign test Binomial test	NA
	Continuous with interval or rational scale	Wilcoxon test	Student *t* test for paired data
To compare two independent samples	Categorical with nominal or ordinal scale	χ^2 test Fisher exact test	NA
	Continuous with interval/rational scale	Mann-Whitney *U* test	Student *t* test for independent data
To compare three or more dependent samples	Categorical with dichotomous scale	Cochran *Q* test	NA
	Categorical with ordinal scale or continuous with interval/rational scale	Friedman test	Two-way ANOVA (*F*-test)
To compare three or more independent samples	Categorical with nominal or ordinal scale	χ^2 test	NA
	Continuous with interval/rational scale	Kruskal-Wallis test	One-way ANOVA (*F*-test)
To estimate the association strength between two variables	Continuous with rational scale	Spearman correlation coefficient	Pearson correlation coefficient

NA = not available.

Generalist test

in the case of the χ^2 test for two independent samples, the null hypothesis takes into account all the characteristics of the corresponding distributions. This is a generalist test: it can be used to assess the overall significance due to differences in the central tendency, data spread, symmetry, etc. It is not a dedicated test for any one of these indices and, if it returns a significant result, other tests need to be used to understand which index gave rise to the difference.

Post-hoc analysis

A similar observation may be made about all the tests which compare three or more samples (dependent or independent): these tests provide an overall result on, for example, the equality of the medians. A significant result allows us to reject the null hypothesis and therefore accept the alternative hypothesis. However, it does not allow us to establish which pairs of samples gave rise to the significance. In these cases, further analysis is required – so called *"post-hoc" analysis*.

A general consideration has to be made regarding the validity of non-parametric tests.

As stated above, non-parametric tests do not require normal data distribution and, for this reason, they are often indiscriminately used. However, even when using non-parametric tests some conditions need to be verified, which despite being less important than those for parametric tests, limit the applicability of these tests for small samples. Note that almost all the tests presented involve requirements on the sample size. Moreover, these tests are only a part of the whole set of available non-parametric tests. Many other tests have been developed to test various hypotheses. However, *"most of the statisticians might survive with a set of about a dozen of tests"* [Greenhalgh, 2006].

A cookbook with a dozen of test

Non-parametric tests are less rebuttable

There is no consensus of opinion regarding the choice between parametric and non-parametric tests. Some authors believe that parametric tests are preferable, even when it cannot be demonstrated that the data have been extracted from a normal distribution. Some others prefer non-parametric tests because, despite being generally less powerful (often with a small difference or at times even more powerful), they are more reliable and therefore less rebuttable. The debate on the more appropriate test has provided no objective universal answers, but only general guidelines. It is here useful to remember that in all unclear cases one may use the two types of tests, because the comparison between their results allows one to obtain more information on the estimated probability [Soliani, 2007].

Table 5.4 shows the criteria for choosing the right test to be performed for various experimental conditions. To be complete we also included the linear regression methods which will be presented in Chapter 6.

References

Altman DG (1991) Practical statistics for medical research. London. Chapman & Hall, pp 210-212

Armitage P, Berry G (1994) Statistical methods in medical research. 3rd edn. Oxford. Blackwell

Boehm T, Willmann JK, Hilfiker PR et al (2003) Thin-section CT of the lung: does electrocardiographic triggering influence diagnosis? Radiology 229:483-491

Caracciolo E (1992) Introduction to the 2nd italian edn. of: Siegel S and Castellan NJ jr. Statistica non parametrica: Milan. Mc-Graw-Hill

Conover WJ (1999) Practical nonparametric statistics. 3rd edn. New York. Wiley
Greenhalgh T (2006) How to read a paper. The basics of evidence-based medicine. 3rd edn. Oxford. BMJ books, Blackwell, p 79
Leisenring W, Pepe MS, Longton G (1997) A marginal regression modelling framework for evaluating medical diagnostic tests. Stat Med 16:1263-1281
Leisenring W, Pepe MS (1998) Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. Biometrics 54:444-452
McNemar Q (1969) Psychological statistics. 4th edn. New York. Wiley
Sardanelli F, Zandrino F, Imperiale A et al (2000) Breast biphasic compression versus standard monophasic compression in x-ray mammography. Radiology 217:576-580
Siegel S, Castellan NJ jr (1992) Statistica non parametrica: 2° edizione italiana a cura di Caracciolo E. Milan. Mc-Graw-Hill
Soliani L (2007) Statistica applicata alla ricerca e alle professioni scientifiche. Manuale di statistica univariata e bivariata. Parma. Uninova-Gruppo Pegaso; 9: 1-2 (http://www.dsa.unipr.it/soliani/soliani.html)
Willmann JK, Weishaupt D, Lachat M et al (2002) Electrocardiographically gated multi-detector row CT for assessment of valvular morphology and calcification in aortic stenosis. Radiology 225:120-128
Yates F (1934) Contingency tables involving small numbers and the test χ^2. J R Stat Society Suppl 1:217-235

6

Linear Correlation and Regression

Error always consists
in making a wrong inference,
that is, in ascribing a given effect
to something that did not cause it.

ARTHUR SCHOPENHAUER

When conducting a radiologic research study on a sample of patients or healthy volunteers, a database to collect the data needs to be built. These data may be of various types (patient history, clinics, radiology, histopathology, etc.). In practice, we obtain the information thought to be useful to the research study for each enrolled individual. Frequently, we are interested in understanding if there are relationships between the data, i.e. in testing the existence of *associations* between the variables.

Association

The quantification of the relationships between variables is done by correlation and regression analyses. Unlike what we stated in the previous chapters, this type of statistical analysis involves the measurement of two or more variables for each statistical unit. In this chapter we will limit ourselves to the simple case of only two variables: *bivariate analysis*. For the general case of more than two variables we refer the reader to specialized texts.

Bivariate analysis

6.1. Association and Causation

Let us consider the following example which will be useful for understanding the way to assess the possible association between two variables.

Example 6.1. MR imaging of prostate for the evaluation of the association between image features and histopathologic Gleason grade. Seventy-four patients undergo endorectal MR imaging before radical prostatectomy to assess the relationship between the signal intensity on T2-weighted images

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

and the histopathologic Gleason grade of the lesion. The authors of this study build a table containing the ratios between the signal intensity of the tumor tissue and that of the obturator muscle and the ratio between signal intensity of the healthy prostate tissue and that of the same muscle. The authors demonstrate a significant ($p = 0.006$) association between the tumor/muscle signal intensity ratio and the Gleason grade for the peripheral part of the prostate (the lower this ratio, the higher the Gleason grade) [Wang et al, 2008].

Association between two variables

An association between two continuous variables means that as one of them increases, the other increases or decreases in value, *although not necessarily involving a cause-effect relationship*. If an association between A and B is demonstrated, then one of the following possibilities can be true:

– A causes B;
– B causes A;
– A and B depend on one or more concomitant factors.

Demonstrating an association is not enough to demonstrate causation

Therefore, we cannot conclude that one of the two variables is the *cause* and the other is the *effect*, because both of them may depend on other factors not taken into account and acting in the *background*. To make the discussion clearer, let us consider the following example.

Example 6.2. Coexistence of stenosis in both cerebral and coronary arteries. Eighty patients with coronary artery disease undergo both coronary and cerebrovascular angiography. The goal of the study is to investigate whether there is an association between cerebrovascular (intra- and extracranial) arterial stenoses and coronary artery stenoses. Considering an artery with a lumen narrowing larger than 50% as stenotic, the authors find only extracranial stenoses in 18 patients (22.5%), only intracranial stenoses in 14 patients (17.5%), and both extracranial and intracranial stenoses in 20 patients (25%). Out of 80 patients, 52 (65%) have coexistence of both coronary and cerebrovascular arterial stenoses ($r = 0.562$, $p < 0.001$) [Li et al, 2007].

In Example 6.2 the authors demonstrated an association between the presence of stenoses of the coronary arteries and the presence of stenoses of the cerebral arteries. The two phenomena have the same pathogenesis and clearly depend on the patient's age. However, we cannot conclude in favor of a causal relationship, i.e. the development of coronary stenoses implies the development of the same disease in cerebral arteries. In this case the third possibility of the list above is at play: over time the same disease develops in different sites. *In the absence of complete understanding of the observed phenomenon, it is always appropriate to speak of association, without coming to a definitive conclusion regarding causal relationship*.

Caution in concluding for a casual relationship

Lastly, if the value of one variable does not influence the value of the other variable, we say that they are *independent* (of each other) variables. In the same study proposed in Example 6.2 [Li et al, 2007], the authors demonstrated that there is no association between the disease degree and the cholesterol blood level: the two variables (stenosis degree and cholesterol blood level) are independent of each other.

In Example 6.1 we reported the association between a continuous variable (the ratio between two values of signal intensities) and an ordinal variable (Gleason grade), but the discussion may be extended to any type of variable. A radiologist might be interested in assessing if in a CT study there is an association between the volume or the contrast enhancement of tumors of the hypophysis or endocrine pancreas secreting hormones and the blood levels of the secreted hormone. In this case, we evaluate two continuous variables.

There are many statistical methods for assessing associations between data and they principally depend on the data type. In this chapter we introduce the main methods.

6.2. Correlation between Continuous Variables

Positive and negative correlation

The statistical technique used for assessing associations between continuous variables is *correlation* analysis. Let x and y be two continuous variables: if in the presence of an increase in the x value we observe an increase in the y value, then there is a *positive correlation*; if in the presence of an increase in the x value we observe a decrease in the y value, then there is a *negative correlation*. One may say that x and y are linearly correlated with each other, or that a *linear correlation* between them exists, when the mathematical relationship is a straight line with the equation:

$$y = ax + b \tag{6.1}$$

In this equation a is the angular coefficient, i.e. a measure of the slope of the line, while b is the intercept, i.e. the intersection point between the straight line and the y-axis. The meaning of a and b may be made clearer by observing Figure 6.1 which shows two straight lines with the same intercept but different angular coefficients.

In medicine, because of wide biological variability, data are rarely perfectly aligned as in Figure 6.1. Data tend to show some spreading about a general trend. Note the following example of positive and negative correlation.

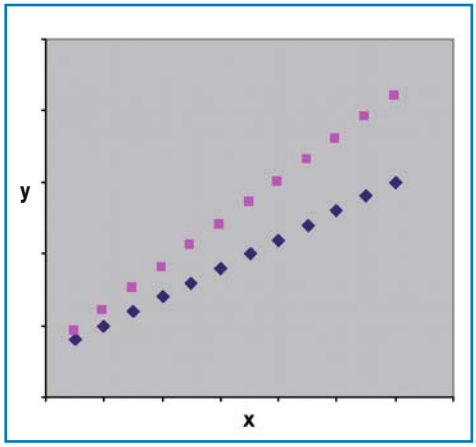


Figure 6.1. The graph shows two straight lines with the same intercept but different slopes.

Example 6.3. Correlation between the uptake of ^{18}FDG and gadopentetate dimeglumine (Gd-DTPA). The authors assess the *in vivo* relationship between the ^{18}FDG (18-fluorodeoxyglucose) uptake at positron emission tomography (PET) and the tumor functional vascularization at MR imaging in patients with colorectal cancer and hepatic metastases. The metastasis metabolism is assessed through the ratio between the uptake of ^{18}FDG by the tumor and that of the liver healthy tissue. From the time course of Gd-DTPA enhancement, the authors calculate the rate constant $k_{ep}(s^{-1})$ of the contrast agent as a measurement of tumor blood flow. Moreover, the vascular density (number of vessels per mm^2 of viable tumor surface) is measured through a computed microscope. The authors demonstrate a negative correlation between the tumor/non tumor (T/NT) ^{18}FDG uptake and the rate constant of Gd-DTPA k_{ep} (Figure 6.2). They also demonstrate a linear positive correlation between Gd-DTPA k_{ep} and vascular density (Figure 6.3). Finally, no correlation between T/NT ^{18}FDG uptake and vascular density is observed ($p = 0.944$) [VAN LAARHOVEN ET AL, 2005].

Pearson correlation coefficient

The correlation is mathematically described by the *correlation coefficient* (or Pearson correlation coefficient), denoted with r. Suppose we measure the variables x and y for a sample of *n* individuals; the linear correlation coefficient is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{n}(x_i - m_x)^2 \sum_{i=1}^{n}(y_i - m_y)^2}}$$

where m_x and m_y are the two arithmetical means.

The correlation coefficient is a measure of the state of being contemporary

This formula may appear complicated, but looking at the numerator we may recognize that it is a measure of the *state of being contemporary* of the variation of x and y. Once the index of the sum, i, is fixed (i.e. once a sample individual is chosen), the differences $(x_i - m_x)$ and $(y_i - m_y)$ indicate the variation of the two variables with respect to their means. The more the differences vary simultaneously from each other, the larger is their product. If there is no rela-

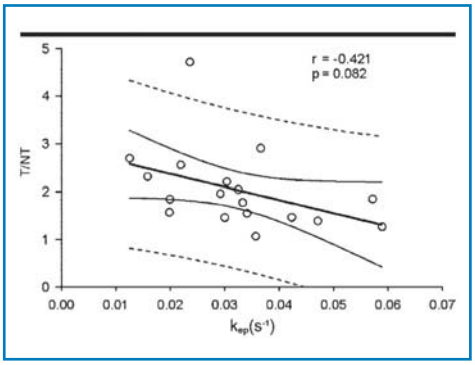


Figure 6.2. The graph shows the relation between the T/NT ^{18}FDG uptake and the rate constant of gadopentetate dimeglumine k_{ep}, as well as the regression line, the corresponding confidence interval, the confidence interval for the single measurement (see next sections), the value of the Pearson correlation coefficient, and the corresponding *p* value. From: van Laarhoven HWM et al (2005) Radiology 237:181-188 (with permission of the authors and of the copyright owner [RSNA]).

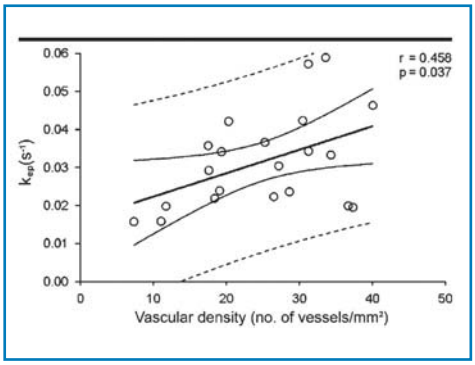


Figure 6.3. The graph shows the relation between the rate constant of gadopentetate dimeglumine k_{ep} and the vascular density, as well as the regression line, the corresponding confidence interval, the confidence interval for the single measurement (see next sections), the value of the Pearson correlation coefficient, and the corresponding *p* value. From: van Laarhoven HWM et al (2005) Radiology 237:181-188 (with permission of the authors and of the copyright owner [RSNA]).

tionship between the two variables, the variations $(x_i - m_x)$ and $(y_i - m_y)$ are completely random and their product on average is zero. Conversely, if an increase in the difference $(x_i - m_x)$ is accompanied by an increase in the difference $(y_i - m_y)$, then the numerator tends to increase, and the more x and y are aligned, the larger r is.

The key information we want to obtain is completely contained in the numerator of the correlation coefficient[1]. The denominator is introduced only to make r a pure (without measurement unit) coefficient, so as to allow the direct comparison of two or more studies[2].

The correlation coefficient may take all the values included in the interval [-1, 1]: positive r values indicate that if x increases (i.e. if we shift from an individual with a certain x value to another with a larger x value), then the corresponding y value increases as well; negative r values indicate the opposite trend. An r value close to zero indicates that there is no linear relationship, although a mathematical relation other than the linear type may exist. An r value equal to 1 or to -1 is observed *only* when the graph points are *perfectly* aligned as they are in Figure 6.1: *the r value indicates the data alignment degree along a straight line*. The better the points of the graph are aligned, the closer r is to 1 or -1, regardless of the line slope.

The correlation coefficient varies between -1 and 1

6.3. Interpreting the Correlation Coefficient

The linear correlation coefficient does not represent the amount of increase (or decrease) of y when x increases, it is rather a measure of the degree of alignment of the experimental points along a straight line. The *association strength* is expressed by the line slope which, in Equation (6.1), is represented by the a coefficient. *Only when by "association strength" we mean the trend for the*

The correlation coefficient is a measure of the degree of alignment between the experimental points

[1] To be rigorous, the numerator is a measure of the *covariance*.
[2] Note that the denominator of the correlation coefficient is equal to $(n - 1)s_x s_y$, i.e. the product of the standard deviation of x (s_x) and of y (s_y) and the number of degree of freedom (n – 1).

experimental points to "associate around" the line may we state that r is a measure of the association strength.

To confirm what we have just stated, the reader may observe Figure 6.4, which reports the data of two independent samples: although the pink points (sample 1) follow the line with the steeper slope, the corresponding r value (0.91) is lower than that the blue points (sample 2; r = 1.00). The larger association strength (meaning the slope) is observed for sample 1, while the alignment is better in sample 2 than in sample 1. As you can see, *only the degree of alignment influences the r value*.

You might like to know what difference there is between two samples whose r values, even if large, substantially differ from one another. Observing Figure 6.4, you may ask why sample 1 is characterized by a larger spread than sample 2 which shows no spread at all[3]. In order to answer this question, we need to remember that the data spread depends on several factors, including measurement errors, intrinsic biological variability and, obviously, x variation. The analysis of variance shows that only a part of the y variation depends on the corresponding x variation. This percentage is expressed by the *determination coefficient* defined as $100r^2$, i.e. by the square of the correlation coefficient multiplied by 100. In the case of Figure 6.4, about 83% of the y variation observed in sample 1 (pink points) is associated with the x variation, while the remaining 17% depends on other factors. In sample 2 (blue points) this proportion is equal to 100%, a limit case where both biological variability and measurement errors are equal to zero. In this extreme situation, an increase in the x value entirely corresponds to a y variation, with a perfectly linear trend.

The determination coefficient

Lastly, the correlation coefficient is dedicated to assessing linear relationships, and even though it may be calculated on data showing curved behavior, it has a no meaning in these cases. In Figure 6.5, r = 0.87 should indicate a high correlation, but the graph points are aligned along a parabolic curve.

The correlation coefficient is aimed to assess linear relationships

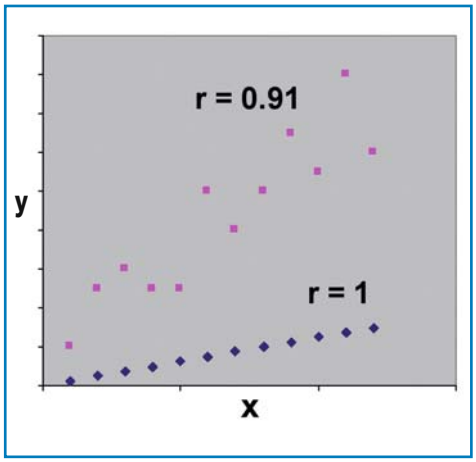


Figure 6.4. Graph showing data of two independent samples with different association strengths. The example shows how the r value depends on the alignment and not on the line slope.

[3] By choice, the data in sample 2 are perfectly aligned to stress the conceptual difference between association strength and alignment.

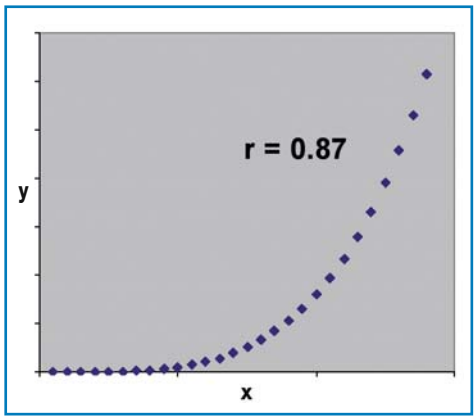


Figura 6.5. Graph showing a set of points with parabolic behavior. The example demonstrates that limiting ourselves to calculating r is not enough and that we need to verify if a linear relationship exists.

6.4. Test for Significance

Consider the following example.

Example 6.4. Relationship between number of excitation and noise in MR imaging. Suppose we assess the relationship between image noise and number of excitation (NEX) using a certain pulse sequence in MR. To do so, ten patients undergo a brain MR imaging examination with an increasing NEX, patient-by-patient[4]. Noise, expressed in arbitrary units (a.u.), is measured in a region of interest placed in a patient- and artifact-free part of the field of view. Table 6.1 shows the results.

Once we have calculated the correlation coefficient, we may ask: "Is the observed r value significant? In other words, are the two variables actually cor-

Table 6.1. Number of excitations (NEX) and noise measurements in MR imaging

Patient	NEX	Noise (a.u.)
1	2.0	10.3
2	2.2	12.1
3	2.6	10.2
4	3.0	10.5
5	3.5	6.7
6	3.6	8.2
7	4.2	8.3
8	4.8	4.2
9	5.0	5.1
10	5.3	3.0
		r = -0.93

a.u.= arbitrary units.

[4] The *phase oversampling* technique enables a fractional NEX value to be selected.

related or is the observed association apparent, probably due to the wide data variability?" An apparently high r value might also be not significant; on the other hand, a low r value may unexpectedly prove significant.

Test for significance of the correlation coefficient

To answer the previous question we need to perform a significance test with the null hypothesis H_0: r = 0, i.e. that the two studied variables are not correlated.

If the null hypothesis is true, then the quantity

$$r\sqrt{\frac{n-2}{1-r^2}}$$

has a Student *t* distribution with n – 2 degrees of freedom. The value calculated with the previous formula has to be compared with suitable published tables [Altman, 1991] to retrieve the corresponding *p* value and to establish the significance of r. For Example 6.4, r = -0.93 provides t = 7.11 and $p < 0.001$: therefore, the negative correlation between the two variables is highly significant.

Now let us consider a simple generalization of the significance test that may be useful in some circumstances. Sometimes it may be interesting to change the null hypothesis to verify that the correlation coefficient is not statistically different from a fixed value r_0. The new hypothesis is H_0: $r = r_0$. This approach is used when the researcher knows that the two variables correlate with each other and wants to verify if the correlation coefficient is equal to or larger than the hypothesized value (r_0). For sample size $n \geq 30$, it may be demonstrated that the standard error of r is approximately equal to:

$$\frac{(1-r^2)}{\sqrt{n}}$$

and that the quantity

$$z = \frac{r - r_0}{(1-r^2)/\sqrt{n}}$$

has an approximated standard normal distribution. The observed z value has to be compared with suitable published tables [Altman, 1991] to retrieve the corresponding *p* value.

6.5. Rank Correlation

The use of the Pearson correlation coefficient depends on some *a priori* assumptions that limit its applicability. The studied variables need to be measured on a random sample and at least one of them should have normal distribution. Preferably, it would be better if both variables have Gaussian distribution, especially when performing the significance test.

The fastest way to verify these hypotheses is to build the histogram of both the variables: it will suffice to check that the two histograms approximately have normal distribution. If not, a valid modification of the Pearson correlation coefficient is its non-parametric version, known as *rank correlation coefficient* or *Spearman correlation coefficient*.

Spearman correlation coefficient

In Table 6.2 we have retrieved the data of NEX and image noise from Example 6.4 and added two columns reporting the corresponding ranks.

The Spearman correlation coefficient, r_s, may be calculated with the formula for the Pearson correlation coefficient using the ranks instead of the original data. Like before, if the null hypothesis H_0: $r_s = 0$ is true, then the quantity:

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

has normal distribution for sample size $n \geq 30$. The observed t value has to be compared with suitable published tables [ALTMAN, 1991][5]. For data reported in Table 6.2, $r_s = -0.88$, $t = 5.22$ and $p < 0.001$. Thus, the significance of the linear relationship between the two variables is confirmed.

Let us consider another example with the use of the rank correlation coefficient.

Example 6.5. Spearman correlation coefficient. Twenty patients with bone marrow lesions undergo spinal MR imaging using spin-echo and four-echo Carr-Purcell-Meiboom-Gill sequences. For each patient, T1 and T2 values are obtained through regions of interest placed in L2, L3, and L4 vertebrae. The bone marrow cellularity is measured by morphometric count techniques. The authors analyze the correlation between T1 or T2 and cellularity, calculating the rank correlation coefficient with the following results: $r_s = 0.74$ (T1 versus cellularity) with $p < 0.001$ and $r_s = -0.18$ (T2 versus cellularity) with $p = 0.1$ [SMITH ET AL, 1989].

Differences between Pearson and Spearman coefficients

The numerical difference between the two correlation coefficients is a measure of the satisfaction degree of the hypotheses we introduced at the beginning of this section: the more the two distributions are different from normal distribution, the larger the difference between the two coefficients is [SOLIANI,

Table 6.2. Data from Example 6.4 with ranks

Patient	NEX	Rank	Noise (a.u.)	Rank
1	2.0	1	10.3	8
2	2.2	2	12.1	10
3	2.6	3	10.2	7
4	3.0	4	10.5	9
5	3.5	5	6.7	4
6	3.6	6	8.2	5
7	4.2	7	8.3	6
8	4.8	8	4.2	2
9	5.0	9	5.1	3
10	5.3	10	3.0	1
			$r_s = -0.88$	

a.u. = arbitrary units.

[5] Obviously, the significance test for both r and r_s is performed by a statistical software package, as with all the other statistical tests.

2007]. In Example 6.4, we obtained r = -0.93 and r_s = -0.88, with a poor difference: in this case we may choose which of the two coefficients to use as correlation coefficient without any other verification. However, when this difference increases, the rank correlation coefficient, which requires no assumption regarding variable distributions, should always be reported.

6.6. Linear Regression

An extension of correlation analysis

In many statistical texts the discussion of correlation is often followed by *linear regression* which, despite being a different statistical technique, shares the basic concepts of correlation. Indeed, linear regression analysis may be considered as an extension of correlation, as it provides more general information on the same data.

When we measure two or more variables for all the individuals of a certain random sample, in addition to checking whether any correlations between the data exist, we are often interested in *describing* these correlations with mathematical formulas which summarize all information. In Example 6.4 we observed a negative correlation between image noise and NEX in MR imaging but we are not able to *predict* the noise level which corresponds to an intermediate NEX value. To do so, we would need to obtain the noise level corresponding to the hypothesized NEX value.

Now let us reconsider Example 6.4 and report the data of Table 6.1 on a Cartesian graph (Fig. 6.6).

Regression line or line of best fit

The graph in Figure 6.6 confirms the negative linear correlation between the two variables, but it does not provide information about the straight line which would *better approximate* the experimental data. What we are searching for is a mathematical method that provides a line that *passes through all graph points, on average*. Such a line is denoted *regression line* or *line of best fit*.

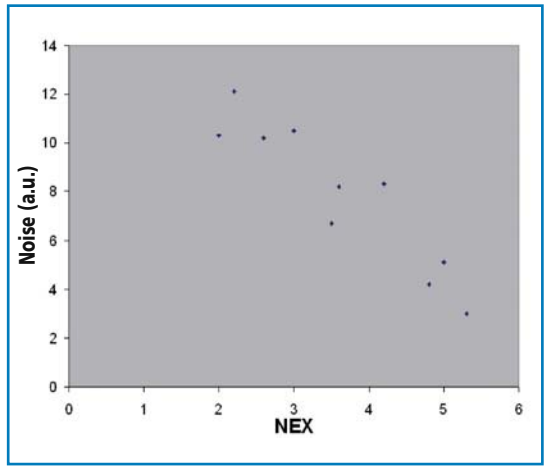


Figure 6.6. Cartesian graph of the data in Example 6.4. The y-axis corresponds to image noise in arbitrary units (a.u.); the x-axis to NEX. The graph shows a strong negative correlation (r = -0.93).

6.6.1. Coefficients for Linear Regression

The least square method

There are many methods for obtaining the regression line. The most used is the *least square method*. This method, whose demonstration is omitted, acts on the quantity

$$\sum_{i=1}^{n}(y^{i}_{observed} - y^{i}_{expected})^2 = \sum_{i=1}^{n}[y^{i}_{observed} - (a \cdot x^{i}_{observed} + b)]^2 = \text{minimum} \quad (6.2)$$

We obtain both the a and b coefficients that minimize the sum. At first glance, this formula might appear complicated, but Figure 6.7 will help the reader to clearly understand the approach of this method.

The vertical bars represent the difference between the $y^{i}_{observed}$ (which corresponds to the $x^{i}_{observed}$) of the i-th sample individual and the expected value based on the regression line ($a \cdot x^{i}_{observed} + b$) relative to the same $x^{i}_{observed}$ value. This difference is called *residual*. In Equation (6.2) all residual squares are summed: the larger the sum, the poorer the *goodness of fit*. *Therefore, the regression line is the line that reduces the sum of the residual squares to a minimum, trying to pass as close as possible to all graph points*. The reader will have recognized that the goodness of fit depends on the degree of alignment of the graph points and, therefore, on the linear correlation coefficient.

Let m_x and m_y be the mean values of the two variables and let x_i and y_i be the single values of the sample. It may be demonstrated that the regression line passes through the point of coordinates (m_x, m_y) and that it is:

$$m_y = am_x + b$$

from which we immediately obtain the intercept as:

$$b = m_y - am_x$$

that may be calculated once we obtain the slope as:

$$a = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sum_{i=1}^{n}(x_i - m_x)^2}$$

The slope calculation may be simplified once we calculate the *sum of squares* and the *sum of products*:

$$S_{xx} = \sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2/n$$

$$S_{yy} = \sum_{i=1}^{n}y_i^2 - (\sum_{i=1}^{n}y_i)^2/n$$

$$S_{xy} = \sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i/n$$

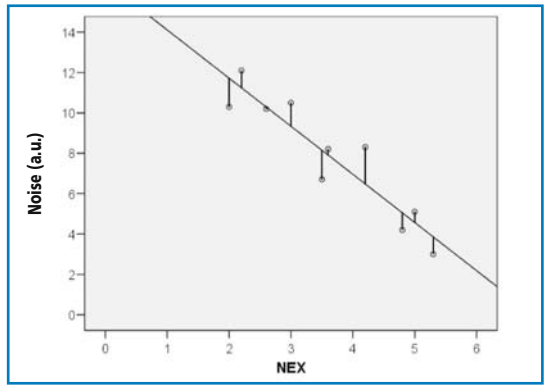


Figure 6.7. Regression line for the data in Example 6.4. The vertical bars indicate the residuals, i.e. the differences between the noise of each point and the expected noise based on the regression line.

In this case:

$$a = \frac{S_{xy}}{S_{xx}}$$

To summarize, if you have a table similar to the one in Example 6.4, where two continuous variables are measured, firstly you should report the data on a Cartesian graph to provide a first impression of the data trend. If this graph suggests a linear relationship, calculate both the Pearson and Spearman coefficients. Secondly, calculate the two arithmetic means and the two regression coefficients, *a* and *b*.

Consider the following example.

Example 6.6. Relationship between muscle fibers and T1 and T2 relaxation times in MR imaging. The authors hypothesize that by measuring the relaxation times T1 and T2 in a region of interest placed in the vastus lateralis muscle it would be possible to assess muscle fiber composition and distinguish between fast-twitch and slow-twitch fibers. For this purpose, 16 volunteer athletes undergo muscle biopsy to establish the fiber composition in terms of fast-twitch fiber percentage (%FTf). About two weeks later, they undergo an MR imaging examination using a 0.22-T magnet of the same muscular region (inversion recovery sequence; TR = 2000 ms; TI = 500 ms; TE = 34 ms). Considering %FTf as a dependent variable and T1 and T2 as independent variables[6], they obtain the following results:

%FTf = 0.66T1 – 172.4 (r = 0.924, $p < 0.01$)
%FTf = 4.9T2 – 81.4 (r = 0.889, $p < 0.01$)

with a highly significant correlation [Kuno et al, 1988].

[6] This notation may be confusing to the reader. Remember that in Cartesian graphs the x-axis variable is called "independent variable" while the y-axis variable is called "dependent variable", without any reference to correlation.

The working hypothesis

Example 6.6 provides food for thought. Firstly, it demonstrates a very important aspect of the research: the *working hypothesis*. Kuno et al. began from an intuition: they suspected that the type of muscle fibers may affect the two main MR physical parameters, T1 and T2. Generally, this is the idea, the starting point of studies which have the demonstration of a possible correlation as an endpoint.

An important limitation

Secondly, this example shows how the predictive use of the regression line has to be limited to the observed data interval. Data analysis showed that T1 ranged from 313 ms to 382 ms while T2 ranged from 22 ms to 33 ms, with the corresponding %FTf values ranging from 25% to 95%. For T1 and T2 values beyond the respective data intervals, the two regression lines may provide nonsense results, such as negative values or values larger than 100%. On the other hand, we are not sure that the relationships between the variables still behave in a linear fashion beyond the observed interval where they may also assume curved behavior. For this reason, it is not advisable to extend the results of a linear regression line beyond the observed data interval.

For the sake of completeness, we reported the mathematical formulas for the calculation of the regression line coefficients, but our advice is to use a statistical software package or to ask a Statistician to do the job.

Automated procedures for coefficients calculation

In order to further clarify the least square method, we now briefly describe the algorithm followed by computers when calculating these coefficients. Once the data are placed in a table similar to the one in Example 6.4, the computer initially assigns two random values to the coefficients a and b and then it calculates the residual square sum as in Equation (6.2). In the second step, it retains b, increases a by a small amount and recalculates the residual square sum: if the new value is lower than the previous one, then the computer again increases a while retaining b, until it reaches a minimum point (which it finds when the sum starts to increase). If in the second step the residual square sum is larger than the previous one, then the computer starts to decrease a by the same small amount as before and recalculates the sum, continuing until it reaches a minimum. When it obtains the value of coefficient a that minimizes the sum, it retains a and repeats the cycle for coefficient b.

Adapting the regression line around the experimental points

This automatic computed procedure is the basis of the least square method and gives the reader an idea of the *adapting process of the regression line around the experimental points*, while searching for the *minimum possible error*. For the data in Example 6.4 the regression line is:

$$\text{noise} = -2.39\text{NEX} + 16.50 \text{ a.u.} \tag{6.3}$$

which describes the behavior of the image noise for any given NEX value. The slope a = -2.39 a.u. represents the decrease in image noise for one unit (1 NEX) of the number of excitations: each unitary increase of NEX involves a decrease of 2.39 a.u. in image noise. The intercept b = 16.50, from a mathematical point of view, indicates the noise level when NEX = 0, i.e. the intersection point between the line in Figure 6.7 and the y-axis. As often happens in medicine, the intercept has no physical meaning, since the value of the x variable (the NEX in Example 6.4) can never be zero: setting NEX to zero would mean not performing the MR sequence at all.

6.7. Interpreting the Regression Line

The regression line can be considered the line joining the mean values of the dependent variable (y) for given values of the independent values (x). Let us reconsider Example 6.4. We can interpret Equation (6.3) as an estimation of the mean noise value for a given NEX value. If, for example, we perform the same MR sequence in *n* patients with NEX = 4, we obtain a mean noise level equal to (-2.39 · 4) + 16.50 = 6.94 a.u.

The confidence interval of the regression line

As with all estimations obtained from a sample, the 95% confidence interval (95%CI) can also be calculated for the regression line. Figure 6.8 shows the regression line for the data in Example 6.4 with the corresponding confidence interval.

The two curves enclosing the regression line represent the constraints within which the *true* regression line of the whole population can be found, with a confidence level equal to 95%. Similarly, once we fix a NEX value, the two curves give both the inferior and the superior boundaries of the confidence interval of the mean noise estimated by the regression line.

What is the procedure for calculating the confidence interval? Let us consider a sample of *n* individuals of which we measure the variables x and y and let m_x be the arithmetic mean of x. Let y_{fit} be the y value estimated by the regression line corresponding to a given x_0 value, that is:

$$y_{fit} = ax_0 + b$$

It is possible to demonstrate that the standard error of y_{fit} is:

$$SE(y_{fit}) = S_{res}\sqrt{\frac{1}{n} + \frac{(x_0 - m_x)^2}{S_{xx}}}$$

where S_{res} is the standard deviation of the residuals, equal to:

$$S_{res} = \sqrt{\frac{(S_{yy} - aS_{xy})}{n - 2}}$$

The 95%CI is:

$$y_{fit} \pm t_{0.975} \cdot SE(y_{fit}) \qquad (6.4)$$

where $t_{0.975}$ is the *t* value corresponding to an area equal to 0.025 in the *t* distribution with *n* – 2 degree of freedom[7]. Once $t_{0.975}$ is obtained from the published tables [Altman, 1991] and allowing x_0 to vary its value, Equation (6.4) provides the confidence interval of the regression line. In Example 6.4, with $x_0 = 4$, we have:

[7] The reader should note that $t_{0.975}$ does not correspond to the $t_{95\%}$ we defined in previous chapters.

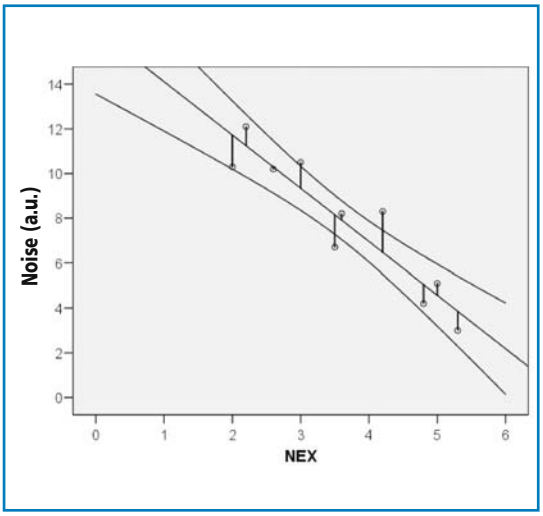


Figure 6.8. Regression line for the data in Example 6.4. The two curves represent the 95% confidence interval, i.e. the constraints within which the *true* regression line of the entire population can be found.

$$a = -2.39 \text{ a.u.}$$
$$b = 16.50 \text{ a.u.}$$
$$x_0 = 4$$
$$y_{fit} = 6.94 \text{ a.u.}$$
$$S_{xx} = 12.54$$
$$S_{yy} = 82.66 \text{ a.u.}^2$$
$$S_{xy} = -29.90 \text{ a.u.}$$
$$S_{res} = 1.18 \text{ a.u.}$$
$$SE(y_{fit}) = 0.40 \text{ a.u.}$$
$$t_{0.975} = 2.31$$
$$95\%CI = 6.94 \pm 2.31 \cdot 0.40 = [6.02, 7.86] \text{ a.u.}$$

The 95%CI indicates that with a 95% confidence level, the image noise acquired with NEX = 4 ranges from 6.02 a.u. to 7.86 a.u. As the reader can observe, the width of the 95%CI is quite small, due to the good alignment of the experimental points of Figure 6.8.

With regard to the confidence interval of the slope a, it may be demonstrated that its standard error is:

$$SE(a) = \frac{S_{res}}{\sqrt{S_{xx}}}$$

The 95%CI of the slope is:

$$a \pm t_{0.975} \cdot SE(a)$$

where $t_{0.975}$ is the *t* value corresponding to an area equal to 0.025 in the *t* distribution with n – 2 degree of freedom. Lastly, we can perform a test for significance with the null hypothesis H_0: a = 0, where the regression line is not significantly different from a line parallel to the x-axis. This hypothesis can

be tested by verifyng whether the correlation between the two variables is statistically significant, i.e. the correlation coefficient is larger than 0. For Example 6.4, the 95%CI of the a coefficient is the interval [-3.16, -1.61], with $p < 0.01$.

6.8. Limitations of the Use of the Regression Line

Limiting the inference to the observed data interval

One of the main limitations of the use of regression analysis is the restriction of the inference toward the whole population only to the observed data interval. We are not authorized to calculate the dependent variable value outside the range used for regression analysis. This concept must be stressed because physical and biological phenomena tend to have curvilinear behavior when going beyond certain boundaries. For example, take the darkness of a film when exposed to x-rays. The graph of the optical density versus the radiation dose absorbed is linear within a certain dose interval, but tends to curve and to reach a saturation level for higher dose values.

Requirements for using linear regression analysis

Moreover, the use of linear regression analysis is subject to the verification of the following hypotheses:

– the values of the dependent variable (y) must have normal distribution for each value of the independent variable (x);
– the standard deviation of y must be the same for each x value;
– the relationship between x and y must be linear.

The latter of these may appear repetitive, but this is necessary. Even though this statistical technique may be applied to each pair of continuous variables, it loses its meaning when applied to data with a curvilinear graph, as already stated for the correlation coefficient. However, unlike the correlation coefficient, when performing regression analysis both variables need not be normally distributed.

References

Altman DG (1991) Practical statistics for medical research. London: Chapman & Hall

Kuno S, Katsuta S, Inouye T et al (1998) Relationship between MR relaxation time and muscle fiber composition. Radiology 169:567-568

Li AH, Chu YT, Yang LH (2007) More coronary artery stenosis, more cerebral artery stenosis? A simultaneous angiographic study discloses their strong correlation. Heart Vessels 22:297-302

Smith SR, Williams CE, Davies JM, Edwards RHT (1989) Bone marrow disorders: characterization with quantitative MR imaging. Radiology 172:805-810

Soliani L (2007) Statistica applicata alla ricerca e alle professioni scientifiche. Manuale di statistica univariata e bivariata. Parma: Uninova-Gruppo Pegaso; 21:1-6

van Laarhoven HWM, de Geus-Oei LF, Wiering B et al (2005) Gadopentetate dimeglumine and FDG uptake in liver metastases of colorectal carcinoma as determined with MR imaging and PET. Radiology 237:181-188

Wang L, Mazaheri Y, Zhang J et al (2008) Assessment of biological aggressiveness of prostate cancer: correlation of MR signal intensity with Gleason grade after radical prostatectomy. Radiology 246:168-176

7

Reproducibility: Intraobserver and Interobserver Variability

Who shall decide when doctors disagree?

ALEXANDER POPE

In clinical practice, the radiologist interprets an examination by qualitative evaluation and/or based on the value of continuous variables such as lymph node diameter, ejection fraction of the two cardiac ventricles, degree of stenosis of an artery, etc. Moreover, in cases of qualitative assessment, her/his judgment may be given either as a dichotomous variable (yes/no) or as an ordinal variable.

In Chapter 1 we introduced the diagnostic performance of an imaging modality compared with a reference standard. This chapter deals with a more general discussion which answers the following question: *What is the degree of the intrinsic reliability of a measured value?* In other words: *If we repeat the same measurement n times, what is the probability of obtaining the same value?*

What is the intrinsic reliability of a measured value?

7.1. Sources of Variability

The result of a measurement, whatever the variable may be (ventricular volume, sensitivity, a proportion), is only an estimation of this variable. In most cases the radiologist only records the first obtained value and rarely repeats the measurement to improve the precision of the estimate.

An estimation of the measured variable

Let us consider the following example. A post-ischemic patient undergoes cardiac MR imaging using a cine sequence for the assessment of left ventricular function. In this case, the radiologist measures the volume of the ventricular cavity both in the systolic and diastolic phase and calculates the ejection fraction. Let us now have a closer look at this measuring process. Figure 7.1 shows the procedure which uses a software package suitable for this type of analysis.

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

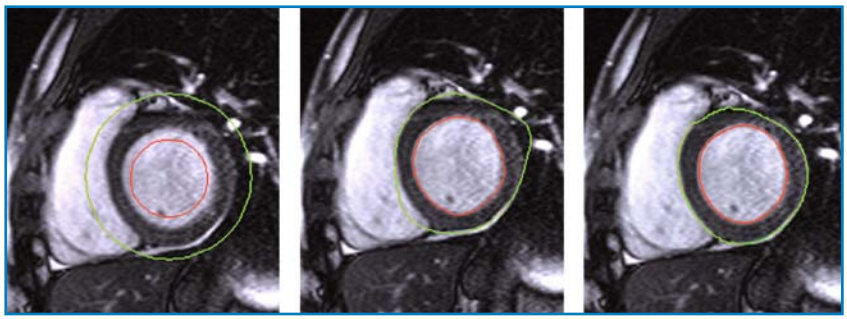


Figure 7.1. The same short-axis end-diastolic image of the heart, obtained at 1.5 T with a four-channel phased-array coil (prospective ECG trigger; true-FISP sequence; TR 45-50 ms; TE 1.5-1.7 ms; FA 65°; slice thickness 8 mm; FOV 196 × 262 mm; matrix size 160 × 256 pixels) is shown with three different phases of segmentation. On the left, two circumferences are placed by the radiologist in such a way that the computer can fit one to the epicardial contour (outer green circle) and the other to the endocardial contour (inner red circle). This fitting is shown in the central panel. On the right, the radiologist has manually corrected the mistakes of automatic fitting of the outer circle.

Variability is intrinsic to the measurement process

The radiologist should depict the endocardial surface in all the slices judged to contain ventricular blood both for the diastolic and the systolic phases; if he/she is interested in measuring heart mass, he/she should also delineate the epicardial contour. As the figure shows, the program *tries* to fit each curve to the relative contour using algorithms which, despite their power, rarely provide optimal results. Therefore, the radiologist must *manually* correct the result proposed by the software. In this latter step (and during the choice of the slices as well as the cardiac phases to be segmented) the *observer*[1] introduces measurement *variability*. Since repeating exactly the same procedure is practically impossible, the repetition of the measurement by the same observer will produce different values.

Intraobserver variability

This example introduces the concept of *intraobserver variability*, i.e. the variability which occurs when the same observer repeats the same measurement under the same conditions two or more times. Even if the observer, images, and tools are all the same, small differences in the choice of slices, regions, and cardiac phases to be segmented provide different results. Since the only weak link in this chain is the observer, the variability we observe in these cases is known as intraobserver variability.

Interobserver variability

Now let us consider the differences which arise when the measurement is not performed by a single observer but by two or more observers. Since each observer has his own intraobserver variability, the overall variability is larger than each single contribution. In this case, we are dealing with *interobserver variability*, i.e. the variability that exists between two or more observers. To clarify the difference between intra- and interobserver variability, let us examine the following example which we will develop in the upcoming sections with the introduction of each new concept.

[1] From now on we will name the operator who performs the measuring process the "observer".

Example 7.1. Intra- and interobserver variability. The authors estimate intra- and interobserver variability in segmenting both left and right cardiac ventricles using a semi-automated segmenting system (ISAM, interactive semi-automated method) and standard manual contouring (MC). Two observers, a radiologist with one-year experience of cardiac MR imaging (R1), and an engineer trained to recognize and segment cardiac cine-MR images (R2), perform four segmenting sessions: two independent sessions for each of them with at least a ten days delay between sessions. The two observers measure the ejection fraction for a sample of $n = 10$ consecutive patients with a wide spectrum of cardiac diseases [SARDANELLI ET AL, 2008]. The results of this study are shown in Table 7.1.

Example 7.1 will enable us in the following sections to make a series of considerations on the importance of the variability estimate. Although variability and reproducibility are complementary concepts (if a measurement has high variability it has low reproducibility and *vice versa*) we prefer to continue the discussion in terms of variability because the statistical techniques we will introduce were developed for estimating variability.

Table 7.1. Ejection fraction of the two cardiac ventricles for ten patients, measured through segmentation of short-axis cine-MR images by two observers (R1, R2) with two different techniques

Left ventricle

	ISAM				MC			
Patient	R1-1	R1-2	R2-1	R2-2	R1-1	R1-2	R2-1	R2-2
1	51.8	55.0	51.3	54.3	55.7	61.7	61.4	57.2
2	56.0	52.5	59.4	59.1	57.7	58.2	56.0	63.6
3	57.8	56.5	66.8	65.8	53.9	58.3	70.2	71.5
4	50.4	70.0	55.4	47.1	70.6	73.6	59.2	55.9
5	15.7	18.7	18.2	14.7	18.3	23.6	18.3	22.3
6	62.2	69.2	68.5	63.5	69.4	68.8	71.1	73.5
7	31.4	29.7	30.1	24.4	23.6	22.1	33.7	30.4
8	61.3	56.6	49.0	49.7	61.4	59.0	47.0	45.7
9	21.1	35.0	31.6	33.1	33.2	31.8	32.2	31.6
10	62.5	71.0	71.5	72.9	70.2	72.0	74.4	70.0

Right ventricle

	ISAM				MC			
Patient	R1-1	R1-2	R2-1	R2-2	R1-1	R1-2	R2-1	R2-2
1	23.8	47.6	25.0	47.6	17.2	47.1	18.1	31.5
2	61.0	46.0	50.0	52.2	46.6	46.1	50.7	46.5
3	76.9	73.9	66.7	65.2	68.0	72.0	65.5	62.6
4	42.2	40.0	51.3	46.1	58.5	54.9	38.4	54.5
5	74.9	68.4	30.6	63.0	67.0	70.1	37.4	59.1
6	72.6	43.3	48.5	61.0	67.1	52.6	69.7	54.9
7	48.0	46.0	44.0	46.3	46.4	46.5	45.8	46.4
8	18.1	14.2	22.9	19.5	18.3	12.8	24.7	21.2
9	37.5	36.7	56.1	53.2	14.2	23.1	43.8	35.6
10	28.7	30.9	54.3	41.3	16.0	33.4	68.1	49.9

R1-1 and R1-2 represent the results of the first and second measure by R1; similarly for R2. All ejection fractions are given as percentages. *ISAM* = interactive semi-automated method; *MC* = manual contouring.

7.2. Why do we Need to Know the Variability of Measurements?

The influence of variability

To understand the importance of knowing the measurement variability, let us consider the following example. A patient with ischemic cardiopathy who undergoes surgical remodeling of the left ventricle repeats the MR examination six months after the intervention in order to assess the efficacy of therapy. The radiologist measures the ejection fraction as 46.1%, larger than the value obtained before surgery (38.8%). The question is: *Is the observed difference (7.3%) a real effect of therapy or is it due to intraobserver variability?* In other words: *If we repeat the measurement once more, is the new value closer to 46.1% or to 38.8%?* On the other hand, the observer who measured the ejection fraction six months after surgery might not be the same observer who performed the measurement before the intervention. If the two observers *disagree* on the choice of systolic and/or diastolic phases, they may also produce very different values of ejection fraction, thus causing the observed difference. This consideration may also be made when the second MR examination is performed using a different MR technique or a different MR unit[2].

The problems presented here raise serious doubts regarding the interpretation of an observed difference. Thus, the key point is: *How should an observed difference in the measured variable be interpreted?* Knowing the variability of measurement is clearly very useful before drawing conclusions. This variability may be expressed in terms of *the least detectable difference*. This parameter is a way of understanding *how large* a difference should be to be considered an effect not due to measurement variability, within a certain confidence level.

The least detectable difference

The reader will have noted the close link with the concept of confidence interval. In fact, one way of answering these questions is to simply compare the confidence intervals of the two estimations or to test the null hypothesis which states that the difference between the two measurements is zero. Let us reconsider the example of the patient who repeated MR imaging six months after surgical cardiac remodeling. A radiologist who performed both ejection fraction measurements (before and six months after the intervention) should assess the efficacy of treatment by comparing the two corresponding confidence intervals.

However, this approach has two important limitations. First, it no longer holds if the measurements are taken by different observers, adding interobserver variability. Second, in clinical practice there is little or no time available for repeating the same measurement, in part due to the need of a suitable interval between measurements to avoid a *learning effect*, i.e. the tendency for an observer to obtain the same results when repeating the same measurement after a short time. Therefore, it is more practical to perform a *preliminary* analysis of intra- and interobserver variability.

A preliminary analysis of the variability enables us to avoid repeating the measurements

Another important aspect needs to be considered. Even if the automatic algorithm of our software does not introduce variability sources[3], it should be born

[2] Examples include the performance difference between 1.5-T and 3-T units, or the use of coils with a different number of channels.

[3] This statement is not precisely true. Some procedures, especially with statistical software, begin with random assignment to temporary variables. On rare occasions the final result of such procedures may depend on the initial random assignment. Note the minimization of χ^2 in Section 6.6.1.

Variability sources sum each other

in mind that when two or more uncertainty sources are present the overall variability is a *weighted sum*[4] of the individual components. Recalling the example of cardiac MR imaging six months after surgery, measurements performed by two different observers on images acquired with two different MR units will be characterized by a variability consisting of the following elements:

1. the intraobserver variability of the radiologist who performed the measurement prior to surgery;
2. the intraobserver variability of the radiologist who performed the measurement after surgery;
3. the interobserver variability;
4. the interstudy variability, due to the repetition of the MR examination;
5. the inter-instrumentation variability, due to the use of two different MR units;
6. the biological variability, due to changes in the patient's health status during the six months elapsed between the two examinations (the effect of therapy may also be a part of this variability).

All these variability sources *act simultaneously* and, as a consequence, the overall variability is a weighted sum of all sources.

In the next sections we will see how to estimate the intra- and interobserver variability for both continuous and categorical variables.

7.3. Intraobserver and Interobserver Variability for Continuous Variables: the Bland-Altman Analysis

John M. Bland and Douglas G. Altman [Bland and Altman, 1986; Bland and Altman, 1999] developed a statistical technique which bears their names for comparing two methods of measurement in medicine. In their studies, the term *method* refers to the instrumentation. Many authors have since extended this approach to the evaluation if intra- and interobserver variability for continuous variables.

When comparing the standard method with a new one, the purpose is to demonstrate that the latter provides results similar to the standard method in such a way as to enable their alternative use. In other cases, the new method is so advantageous in terms of invasiveness and/or costs that, despite being less reproducible than the old method, it might replace it anyway. Just how less reproducible it may be is a clinical and not statistical issue and, in the end, this depends on its effects on patient management.

The Bland-Altman analysis results in a value expressed with the same measurement units as the measured variable. This allows for a direct interpretation. The analysis of intra- and interobserver variability may be performed in a parallel way, through a measurement protocol similar to the one in Example 7.1. *It is enough that two observers perform two measurements for each individual*

[4] The way the several elements sum each other is beyond the aims of this book and would make no essential contribution to the discussion.

The Bland-Altman analysis interprets variability in terms of agreement

of the sample[5]. The Bland-Altman analysis interprets the interobserver variability in terms of *agreement* between the two observers: the higher the agreement, the lower the variability. Similarly, the higher the agreement that a single observer has *with himself* is, the lower the intraobserver variability is.

When measuring a variable in this methodologic setting, instead of the true value we are interested in knowing if the measurement is *reproducible*, i.e. if repeating the same measurement in the same conditions we obtain values very close to each other. Let us suppose, for example, that the most commonly used MR sequence for measuring T1 relaxation time provides a value that is systematically 10% lower than the true value. If regardless of this the procedure is the standard one, it will be enough to take this systematic error into account when using those measurements.

A bit of history

For a long time, linear regression analysis has been used to estimate intra- and interobserver variability, reporting the Pearson correlation coefficient as a measure of data agreement. Since Bland and Altman published their article in The Lancet [Bland and Altman, 1986] the use of their method has become very widespread. We will now examine the criticisms regarding the use of linear correlation analysis for the evaluation of intra- and interobserver variability.

Criticisms regarding the use of linear correlation analysis

Now we will try to evaluate the intraobserver variability of the first observer (R1) in Example 7.1 when measuring the ejection fraction of the left ventricle using the ISAM method. Figure 7.2 reports the second measurement (R1-2) versus the first measurement (R1-1) of the same reader. As you can see, the r value is quite large, indicating a high correlation among the data. However, the two Cartesian axes report the same variable, i.e. the ejection fraction in the same patients at the same time: *clearly there is a correlation, the opposite would be hard to believe*. In actual fact, we are not assessing the possible correlation

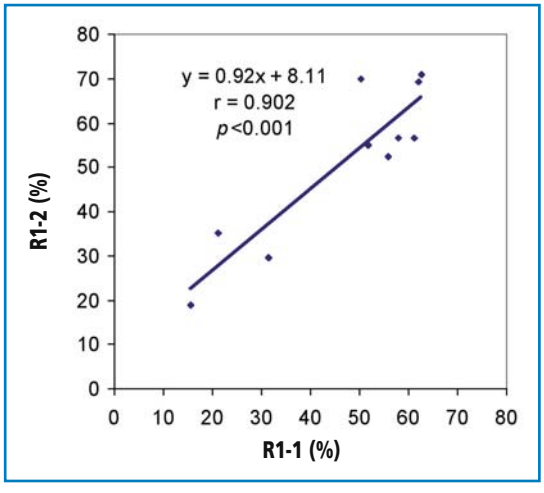


Figure 7.2. This Cartesian graph shows the ejection fraction for the ten patients in Example 7.1. The y-axis and the x-axis report the second (R1-2) and the first (R1-1) measurement of the first observer (R1), respectively. The graph also shows the regression line and its equation, the r value and the corresponding *p* value.

[5] Actually, not only is it enough but it is also the only way: *two* measurements for both of the *two* observers.

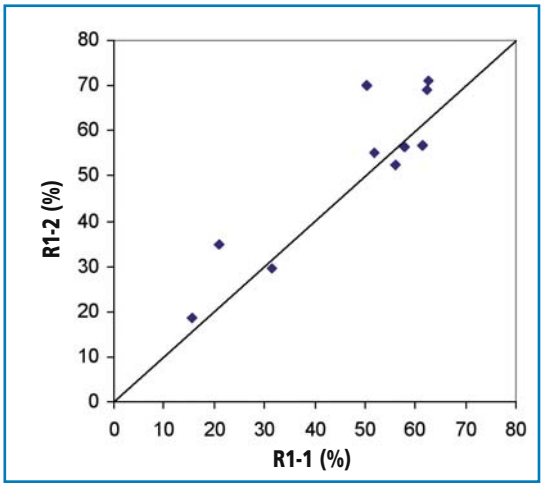


Figure 7.3. This Cartesian graph shows the same data as in Figure 7.2. Note the equality line on which the experimental points would lie in the event of perfect agreement between the first and second measurement.

between two different variables such as the size of a tumor and the blood level of a tumor blood marker, for which there may or may not be an association. The value $p < 0.001$ indicates the probability of the true r value being zero, or rather there is no correlation between the data. However, to estimate the variability, what we need to verify is whether the experimental points lie close to the *equality line*, i.e. the line whose points have identical coordinates.

Figure 7.3 reports the same graph as in Figure 7.2, but instead of the regression line, the equality line is depicted, i.e. the line we would expect if both the first and the second ejection fraction measurements coincide with each other for each patient. This would be the case of perfect agreement. The further the points are from this ideal line, the lower the agreement is. *The spread of the points around this line is a measure of R1 intraobserver variability*.

From a mathematical point of view, the equation of the equality line is[6] $y = x$, i.e. a line with slope $a = 1$ and intercept $b = 0$. If we wanted to use linear regression analysis to estimate the agreement between the two measurements by the same observer, we would verify that the regression line has those coefficients.

By simply observing the graph we have no clear indication of the agreement. The starting point of the method proposed by Bland and Altman is the calculation of the difference between R1-1 and R1-2 for each patient in the sample. In this way we get a new sample made up of these differences, as shown in Table 7.2.

Bias is a mean systematic error

Table 7.2 also reports the mean and the standard deviation of the differences and there is an extra column with the mean of the two measured values. *The mean difference (-4.4% in the example), also called bias, represents the mean error, namely the average quantity that the second measurement adds to or subtracts from the first one*. In practice, the second ejection fraction measurement is 4.4% greater then the first, on average. It is important to take into

[6] We are using the same notation we introduced in Chapter 6, for which the linear equation is $y = ax + b$.

Table 7.2. Application of the Bland-Altman method to the data of Table 7.1

Patient	R1-1 (%)	R1-2 (%)	R1-1 – R1-2 (%)	Mean (R1-1, R1-2) (%)
1	51.8	55.0	-3.2	53.4
2	56.0	52.5	3.5	54.3
3	57.8	56.5	1.3	57.2
4	50.4	70.0	-19.6	60.2
5	15.7	18.7	-3.0	17.2
6	62.2	69.2	-7.0	65.7
7	31.4	29.7	1.7	30.6
8	61.3	56.6	4.7	59.0
9	21.1	35.0	-13.9	28.1
10	62.5	71.0	-8.5	66.8
Mean			-4.4	
Standard Deviation			7.9	

R1-1 and R1-2 represent the first and the second measurement by R1, respectively; similarly for R2.

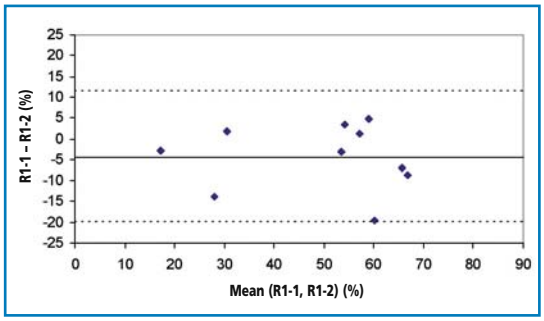


Figure 7.4. Bland-Altman graph for the data in Example 7.1. The y-axis reports the difference between the two measurements (R1-1 – R1-2), while their mean is reported on the x-axis. The continuous line represents the bias, i.e. the mean of the differences, while the dotted lines indicate the limits of agreement (bias ± 2SD). As may be seen, the points are not well centered around zero.

account the order with which we calculate the differences: if we had considered the difference between R1-2 and R1-1, the bias would change its sign.

The distribution of the differences is almost always normal

The variable we are measuring might have non-normal distribution, especially if the sample is extracted from a pathologic population. However, this does not prevent us from using the Bland-Altman analysis. It is more important for the distribution of the differences to be normal, as it generally is due to randomness. This hypothesis enables us to state that 95% of the differences lie in the interval bias ± 1.96SD, where SD is the standard deviation of the differences. This interval defines the *limits of agreement*, often approximated to bias ± 2SD. For Example 7.1, this interval is [-20.2, 11.4]%. Later on we will see how to interpret this interval.

Limits of agreement

Bland-Altman graph

An important feature of the Bland-Altman analysis is the building of the graph bearing their name. Instead of using the graph in Figure 7.3, we need a graph which shows the difference between the two measured values versus their mean, i.e. the last two columns in Table 7.2. Figure 7.4 shows the Bland-Altman graph for the measurement by R1 of the ejection fraction of the left ventricle using the ISAM method (Example 7.1).

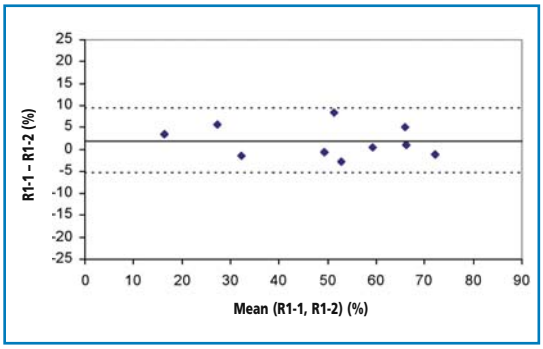


Figure 7.5. Bland-Altman graph for the data in Example 7.1. The y-axis reports the difference between the two measurements (R2-1 – R2-2), while their mean is reported on the x-axis. The continuous line represents the bias, i.e. the mean of the differences, while the dotted lines indicate the limits of agreement (bias ± 2SD).

In the ideal situation, the experimental points would be aligned along the zero line (i.e. the x-axis), which is the same as what we stated above regarding the equality line. Generally, the points will show two features: a more-or-less wide spread and a shift up or down of a quantity that is just the bias. Moreover, the graph shows three horizontal lines: one corresponds to bias (continuous line) and two correspond to the limits of agreement (dotted lines). Based on the main feature of Gaussian distribution, 95% of the points lie within the limits of agreement, on average.

No reference standard

An important point concerns the need to report the mean of the two measurements on the x-axis instead of just one of them. For example, when assessing interobserver variability the senior observer might be thought of as a kind of reference standard. This is an incorrect approach, since the difference between the two measurements is obviously related to the value from which it is calculated and reporting this difference against one of the two measurement introduces a well known statistical artifact. The true value of the variable we are measuring is not known and its best estimation is the mean of the two measurements.

Let us continue to analyze data from Example 7.1. We now take into consideration the R2 intraobserver variability in segmenting the left ventricle using the ISAM method. Repeating what we stated for R1, we find: bias = 1.7%, SD = 3.7% and limits of agreement [-5.6, 9.1]%. Intuitively, the R2 intraobserver variability is lower than the R1. In fact, both the bias and the standard deviation are lower than the values relative to R1 and the limits of agreement have a lower width. Figure 7.5 shows the corresponding Bland-Altman graph. The y-axis scale is the same as before to provide the reader with a visual inspection of the lower variability associated with R2.

Now let us evaluate the interobserver variability between R1 and R2 in measuring the ejection fraction of the left ventricle using the ISAM method (Example 7.1). The reader may wonder which of the two measurements (R1-1 and R1-2 for the first observer, R2-1 and R2-2 for the second observer) to use for the analysis. One of the possibilities is to use their means. However, this choice increases the estimation precision[7] and would result in an underestimation of interobserver variability. A good alternative is to use the first measure-

[7] Remember how the confidence interval of a continuous variable is calculated.

Table 7.3. Results of the Bland-Altman analysis for Example 7.1

	Left ventricle		
	Bias (%)	**SD (%)**	**Limits of agreement (%)**
Variability (ISAM)			
Intraobserver (R1)	-4.4	7.9	[-20.2, 11.4]
Intraobserver (R2)	1.7	3.7	[-5.6, 9.1]
Interobserver	-3.2	6.7	[-16.6, 10.3]
Variability (MC)			
Intraobserver (R1)	-1.5	3.0	[-7.6, 4.6]
Intraobserver (R2)	0.2	4.0	[-7.8, 8.1]
Interobserver	-0.9	9.1	[-19.2, 17.4]
	Right ventricle		
	Bias (%)	**SD (%)**	**Limits of agreement (%)**
Variability (ISAM)			
Intraobserver (R1)	3.7	13.3	[-23.0, 30.3]
Intraobserver (R2)	-4.6	13.9	[-32.4, 23.2]
Interobserver	3.4	20.5	[-37.5, 44.4]
Variability (MC)			
Intraobserver (R1)	-3.9	12.5	[-28.9, 21.0]
Intraobserver (R2)	0.0	13.2	[-26.4, 26.4]
Interobserver	-4.3	23.0	[-50.3, 41.7]

Bias = mean of the differences; SD = standard deviation; Limits of agreement = bias ± 2SD.

ment of both observers (R1-1 and R2-1) since, generally, in clinical practice only one value is measured. In this way we obtain: bias = -3.2%, SD = 6.7% and limits of agreement [-16.6, 10.3]%.

The data analysis of Example 7.1 should also take into consideration the segmentation with the MC method and that of the right ventricle. Table 7.3 shows the final results in terms of bias, SD and limits of agreement.

In the next section we will explain how to interpret these results. Here we limit the discussion to noting that the standard deviation of the differences is systematically larger for the right ventricle than the left ventricle; similarly, the limits of agreement are systematically wider for the right ventricle than for the left ventricle. On the other hand, this result was expected, due to the more complex geometry and the less regular morphology of the right ventricle with respect to the left ventricle. Moreover, by using short-axis images, we end up segmenting the right ventricle on images which are spatially oriented perpendicular to the long axis of the left ventricle, which could not be the best approach for the evaluation of the right ventricle.

7.4. Interpreting the Results of Bland-Altman Analysis

In the previous section we explained how to estimate the intra- and interobserver variability through the Bland-Altman analysis. When presenting the results of this analysis, bias and limits of agreement must be reported.

Now we will discuss the interpretation of these results. Let us consider the interobserver variability between R1 and R2 in measuring the ejection fraction using the manual contouring method of Example 7.1. Suppose that during the segmentation process R2 excludes the papillary muscles to the ventricular cavity. Since the papillary muscles tend to be more visible in the diastolic phase than the systolic phase, there will be a tendency for R2 to measure lower volumes than R1 with a consequent underestimation of the ejection fraction. Therefore, instead of oscillating between positive and negative values around zero, the differences will tend to have positive values, on average[8]. In the Bland-Altman graph the experimental points will not be centered around zero but around a positive value (*barycenter*): in practice the points will be shifted up, on average. The barycenter is the mean of the differences (bias) which is depicted in the graph with a continuous line. *Therefore, the bias represents a systematic error, i.e. the trend for one of the two observer to underestimate or overestimate (with respect to the other observer) the measured variable*.

Bias as a systematic error

In Example 7.1, the R2 intraobserver variability in segmenting the right ventricle using the ISAM method has the following values: bias -4.6%, SD 13.9% and limits of agreement [-32.4, 23.2]%. Since 95% of measurements with a normal distribution lie within the interval mean ± 2SD, the R2 variability is such that if the first measurement had given an ejection fraction value equal to 28.2%, the second measurement could vary from 28.2% – 23.2% = 5.0% to 28.2% + 32.4 % = 60.6%. In other words, the difference between the first and the second measurement may take a negative value (first < second) up to 32.4% and a positive value (first > second) up to 23.2%. This is a very wide interval indeed!

Interpreting the limits of agreement

The coefficient 2SD, i.e. twice the standard deviation of the differences, is also called *coefficient of repeatability*. If, for example, we compare the value of a variable obtained before treatment and that obtained six months after treatment, we have to take into account that differences lower than the coefficient of repeatability cannot be attributed to the treatment, but to chance. *This coefficient assumes the meaning of the least detectable difference: an observed difference should be at least as large as the coefficient of repeatability to be considered real*.

The coefficient of repeatability as the least detectable difference

The repeatability coefficient, 2SD, has the same measurement units as the variable we are measuring. If R1 and R2 in Example 7.1 had measured the ventricular volume (expressed in mL) instead of the ejection fraction, then 2SD should represent the least detectable difference in mL. The result of a Bland-Altman analysis is in itself a continuous variable and it provides more information than a result such as "*The reproducibility is equal to 87%*". Such a result, despite appearing very informative, does not help us to definitively interpret the observed differences.

Let us make another observation. We stated that the limits of agreement are calculated as bias ± 2SD, a formula close to that for the confidence interval calculation. *Be careful of not confusing the limits of agreement and the confidence interval of the bias*.

Limits of agreement are not a confidence interval

Lastly, the reader should note that studies assessing the intra- and the interobserver variability do not necessarily demonstrate that the variability is low and

[8] If the difference is calculated as R1-R2.

Variability may be reduced, not eliminated

that the reproducibility is high. The variability is an aspect intrinsic to the measurement processes. *It may be reduced but not completely eliminated*. A way of reducing variability is to make the measurement as objective as possible, by defining rules and procedures for carrying it out. In Example 7.1, a method for reducing interobserver variability could be the definition of a common protocol in selecting the slices and the phases to be segmented and in segmenting images (e.g. inclusion or exclusion of the papillary muscles). An error to be avoided is to select the patients of the sample from among those with the best images. *The variability exists in itself* and all we have to do is estimate it. Therefore, we should choose a sample which is representative of clinical practice.

7.5. Intra- and Interobserver Variability for Categorical Variables: the Cohen k

Until now we have considered intra- and interobserver variability for continuous variables. In this section we introduce the methods for estimating variability for categorical variables. Unlike continuous variables, the values that a categorical variable may take are often – but not always – the fruit of a personal judgment of the radiologist. An exception is the *discretization* of a continuous variable into two or more categories based on its numerical value. An example is the NASCET criteria [NASCET, 1991] which subdivides the degree of stenosis of carotid arteries (a continuous variable that may take values in the interval [0, 100]%) into the classes *mild* [0, 29]%, *moderate* [30, 69]%, and *severe* [70, 100]%[9]. Notice that the information is greater with continuous variables than with categorical variables and to analyze data based on the original continuous variable is undoubtedly better.

We need a method which provides information on the reproducibility of the judgment of two or more radiologists or the reproducibility a radiologist has with her himself if he/she repeats the evaluation. The logical approach is the same as the one developed in the two previous sections, with the only exception being that now we are studying categorical variables. Unlike the Bland-Altman analysis, the method we are going to introduce expresses the intra- and interobserver variability in terms of reproducibility and provides results in term of a percentage.

We begin our discussion with the data of the following example.

Example 7.2. The Cohen k. Two radiologists, R1 and R2, independently provide a dichotomous judgment (positive/negative) for the presence of secondary hepatic lesions in a sample of 150 abdominal CT examinations. Data are reported in Table 7.4.

Overall agreement

Consider the *concordances*: the number of individuals judged as positive or negative *by both observers* is equal to 7 and 121, respectively. These data are placed on the main diagonal of Table 7.4, while on the secondary diagonal the

[9] The Cohen k is also divided into classes.

Table 7.4. Contingency table of the data in Example 7.2

		R2		
		Positive	Negative	Total
R1	Positive	7	10	17
	Negative	12	121	133
	Total	19	131	150

discordances are placed. The number of individuals judged as negative by R2 and as positive by R1 is equal to 10, while the number of individuals judged as positive by R2 and as negative by R1 is equal to 12. Intuitively, we consider the proportion of the concordances on the total number of patients:

$$p_0 = \frac{7+121}{150} = 0.85$$

This ratio (0.85) represents the overall agreement and is generally expressed as a percentage. Therefore, for the data in Table 7.4, R1 and R2 agree in 85% of their evaluations.

The overall agreement depends on disease prevalence

The overall agreement is rarely used since it does not provide information on the type of agreement. If, as is the case in Example 7.2, one of the components prevails over the others (the number of negative individuals is much larger than the number of positive individuals), p_0 may give a false feeling of high performance. Let us take the example of mammographic screening: since most women have a negative exam, the probability that both observers give a negative judgment is very high and this hides the possible agreement or disagreement on the positive cases.

A good alternative is to separately calculate the agreement on the positive cases, p_+, and on the negative cases, p_-. For Example 7.2:

$$p_+ = \frac{7+7}{(10+7)+(12+7)} = 0.39$$

In practice, the positive concordances (both the observers judge 7 cases as positive) are added and this sum is divided by the total number of cases judged as positive (19 by R2 and 17 by R1). Similarly:

$$p_- = \frac{121+121}{(10+121)+(12+121)} = 0.92$$

the negative concordances (both the observers judges 121 cases as negative) are added and this sum is divided by the total number of cases judged as negative (131 by R2 and 133 by R1).

As you can see, the overall agreement is between p_+ and p_- and is strongly affected by p_-. In actual fact, R1 and R2 highly agree only on the negative cases (92%), while their agreement on the positive cases (39%) is much lower. If we calculate p_+ and p_-, separately any imbalance in the positive/negative propor-

tion clearly appears. The disadvantage is that we cannot calculate the corresponding confidence intervals.

The random agreement

Let us now take a step forward. In addition to disease prevalence and operator experience, we have to consider the probability that the two judgments agree by chance. If for example R1 and R2 were to toss a coin and to judge an examination as positive if they get heads and negative if they get tails, there would be a fraction of patients with identical results. We have to account for this fraction and subtract it from the overall agreement, thus obtaining the *true agreement*. Let p_a be the expected agreement by chance: the true agreement is $p_0 - p_a$. Now we should divide this value by the maximum achievable true agreement $(1-p_a)$.

The Cohen k

In 1960 Jacob Cohen from New York University [COHEN, 1960] proposed a coefficient (later denoted as *Cohen k*) defined as:

$$k = \frac{p_0 - p_a}{1 - p_a}$$

Therefore, the Cohen k is the ratio between the true agreement $(p_0 - p_a)$ and the maximum achievable true agreement $(1 - p_a)$. It is the fraction of the observed agreement on its maximum value not due to chance.

Now we will see how to calculate the expected agreement p_a. To simplify this calculation, we modify Table 7.4 by dividing each cell by the total number of individuals (150 in Example 7.2). Table 7.5 shows the frequency of each cell and it also shows the expected value *calculated as the product of the relative marginal totals*. For the upper left cell the observed frequency is 7/150 = 0.04, while the expected frequency is $0.11 \times 0.12 = 0.01$; similarly for all other cells. The observed overall agreement is the sum of the frequencies on the main diagonal:

$$p_0 = 0.04 + 0.81 = 0.85$$

while the expected agreement is the sum of the expected frequencies on the main diagonal:

$$p_a = 0.01 + 0.78 = 0.79$$

Table 7.5. Contingency table for calculating the Cohen k

		R2		
		Positive	Negative	Total
R1	Positive	0.04 *(0.01)*	0.07 *(0.10)*	0.11
	Negative	0.08 *(0.11)*	0.81 *(0.78)*	0.89
	Total	0.12	0.88	1.00

The expected frequencies are shown in parenthesis.

As the reader may see, 79% agreement was expected simply due to chance, and the remaining true agreement is $p_0 - p_a = 0.85 - 0.79 = 0.06$ (6%), with respect to the maximum value equal to $1 - 0.79 = 0.21$ (21%). The Cohen k is:

$$k = \frac{0.85 - 0.79}{1 - 0.79} = 0.31$$

Thus, we shift from an overall agreement equal to 85% to an *agreement corrected for the effect of chance* (the Cohen k) equal to 31%.

Cohen k ranges between -1 and 1

From a mathematical point of view, the Cohen k may vary in the interval [-1, 1], but the only logical part of this interval is the positive one, i.e. [0, 1]. Many statisticians agree in considering k = 0 as total absence of agreement between the observers and k = 1 as perfect agreement. In 1977, J.R. Landis and G.G. Koch [Landis and Koch, 1977] proposed the classification of the k value shown in Table 7.6.

A paradox?

This classification is arbitrary but it is commonly used. According to Table 7.6, the agreement other than due to chance shown by R1 and R2 in Example 7.2 is fair (31%), though associated with 85% overall agreement. This apparent paradox (large overall agreement and low Cohen k) is due to the high prevalence of negative cases and, therefore, to a very unbalanced data distribution. A more balanced distribution among negative and positive cases would result in a larger value of the Cohen k.

The Cohen k is a measure of the reliability of the overall agreement

A. Feinstein and D. Cicchetti [Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990] explored this paradox suggesting that in studies dealing with intra- and interobserver reproducibility for categorical variables the Cohen k, p_+, and p_- should be reported. In this view, *Cohen k may be considered as a measure of the reliability of the overall agreement*.

Clearly, we cannot trust the overall agreement because it is overly affected by disease prevalence. However, we may assess it by calculating the Cohen k: the closer k is to 1, the more reliable the p_0 value is, whereas the closer it is to 0, the less reliable the p_0 value becomes. Lastly, this approach is similar to the one used for the confidence interval: the reliability of the estimated value depends on the width of the corresponding confidence interval.

The generalized Cohen k

The Cohen k may be generalized to take into account non dichotomous categorical variables and ordinal variables. Data must be organized in a contingency table with *n* rows and *n* columns, where *n* is the number of values that

Table 7.6. Classification of the agreement based on the k value

k	Agreement other than chance
< 0	None
0-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

From: Landis and Koch, 1977.

the variable may take. If two radiologists evaluate a series of mammographies using BI-RADS® 1-to-5 scale, the data should be placed in a 5 × 5 contingency table.

With this generalization, k statistics shows an intrinsic limitation: it only takes into account the concordances and discordances without considering the weight of each discordance. A BI-RADS® discordance between scores 1 (negative) and 2 (benign finding) is clearly much less clinically relevant than the discordance between scores 3 (probably benign finding) and 4 (suspicious abnormality). We can overcome this limitation of k statistics (up to now considered in its simplest form, denoted as *unweighted*) by introducing some coefficients that give different weights to the discordances according to the magnitude of the discrepancy. These coefficients account for the discordances by giving more weight to those discordances the radiologist holds to be more important.

The weighted k

Although we have not provided a mathematical explanation, the reader may easily see that the value of the *weighted k* numerically depends on the chosen coefficients. The arbitrary choice of these coefficients makes the weighted k observer-dependent and does not allow comparison between different experiments. The only way to compare several k values is to use standard weights. However, in medicine, alongside purely statistical considerations we also need to add biological, patient-specific and ethical considerations such as the difference in terms of importance between BI-RADS® scores 3 and 4.

References

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 1:307-310

Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. Statistical Methods 8:135-160

Cicchetti DV, Feinstein AR (1990) High agreement but low kappa. II. Resolving the paradoxes. J Clin Epidemiol 43:551-558

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37-46

Feinstein AR, Cicchetti DV (1990) High agreement but low kappa. I. The problem of two paradoxes. J Clin Epidemiol 43:543-549

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159-174

North American Symptomatic Carotid Endarterectomy Trial Collaborators (1991) Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Eng J Med 325:445-453

Sardanelli F, Quarenghi M, Di Leo G et al (2008) Segmentation of cardiac cine MR images of left and right ventricles: interactive semi-automated methods and manual contouring by two readers with different education and experience. J Magn Reson Imaging 27:785-792

8

Study Design, Systematic Reviews and Levels of Evidence

> It is a capital mistake to theorize before one has data.
> Insensibly one begins to twist facts to suit theories,
> instead of theories to suit facts.
>
> SHERLOCK HOLMES
> (ARTHUR CONAN DOYLE)

Is the difference due to the variability?

In Section 3.1 we stated that when we observe a difference between two groups or two samples, the first thing we should exclude is that this difference is simply due to the effect of variability within the population from which the two samples were taken. From this we derived the method that the use of probability allows us to reject the null hypothesis (H_0) and to accept the experimental hypothesis (H_1). Therefore, if we have excluded variability within the population, does this mean we have a direct demonstration of the experimental hypothesis? Unfortunately, this is not the case. Before we can conclude in favor of the experimental hypothesis, we need to be sure that the entire process (from study design to its practical implementation, in all its details) is free from *bias*, i.e. systematic distortions, which might have influenced the results. If a study is flawed by substantial bias, its application to clinical practice is doubtful or not possible at all.

As with other fields, *prevention is better than cure*. Correctly designing and implementing a study is the right way for producing a good scientific work. In this chapter we will discuss the topic of *study design*, beginning with the classic four phases of pharmacologic research. We will also briefly examine *systematic reviews*, namely those studies which evaluate – using *meta-analysis* – the evidence from already published studies, as well as the hierarchy of the *levels of evidence*, which depend in particular on study design. In the next chapter we will describe the errors to be avoided in studies on diagnostic performance, i.e. the sources of bias in radiologic studies.

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

8.1. Phases 1, 2, 3, and 4 of Pharmacologic Research

Radiologic research on contrast materials

Diagnostic imaging to demonstrate drug activity

At least ten years are needed from the time the manufacturer has the possibly active molecule in their hands to the launch of the new drug onto the market and its use in clinical practice. Radiologic research is obviously involved in the development of contrast materials, but medical imaging is being increasingly used in drug research in a broader sense. In fact, diagnostic imaging can give end-points which are alternative to the clinical course of a disease. These end-points are commonly earlier and more objective than those of the clinical course. This is the case of imaging techniques applied to the study of the number and size of tumors before/after chemotherapy. Criteria for this evaluation were first proposed by the World Health Organization [WHO, 1979]. In 2000, the *response evaluation criteria in solid tumors* (RECIST) appeared as a standardized method [Therasse et al, 2000]. Currently this evaluation is under investigation and the criteria will probably be modified [Therasse et al, 2006]. Fields of clinical research other than oncology use end-points frequently taken from longitudinal imaging studies. An example is the use of MR imaging for the study of new therapeutic approaches for multiple sclerosis [Filippi et al, 1999].

Pharmacodynamics

The main advantage of diagnostic imaging is the possibility of an early non-invasive quantification of the action of an experimental drug, i.e. to provide information on the drug *pharmacodynamics*. In an early experimental phase, imaging modalities give objective, robust, and repeated measurements on small samples of patients, reducing the costs and duration of the studies. In this way, initial promising results from small studies can subsequently be tested in studies with larger samples and an evaluation can be made of drug efficacy in terms of better clinical course, increased years of life adjusted for quality of life and, when applicable, mortality reduction. *Notwithstanding the increasing role of diagnostic imaging in clinical research, we must always remember that the ultimate purpose is to cure the patients, not their images*.

Cure the patients, not their images

Four phases of pharmacologic research

After preclinical cell or animal testing, a new drug reaches the clinical experimentation stage, i.e. human testing. This test on humans is classically subdivided into four phases [Bacchieri and Della Cioppa, 2004; Hoffman et al, 2007] (Table 8.1), with some special features for anticancer drugs. In fact, in oncology, each of the four phases has a particular role:

Peculiar features of clinical research in oncology

- phase 1 includes patients who cannot benefit from other therapies (these patients are frequently in advanced stages of the disease) with the main purpose of finding the right dosage for phase 2 studies;
- phase 2 is aimed at confirming pharmacodynamic action, at least in terms of partial or complete response;
- phase 3 is aimed at demonstrating clinical efficacy in terms of survival rate (and also at verifying safety and tolerability);
- phase 4 involves surveillance after the drug has entered the market.

Pharmacologic studies rarely involve radiologists as principal investigators, with the obvious exception of contrast materials. However, radiologists should have a broader knowledge of the clinical experimentation of drugs, firstly because the fundamental methodology is the same for both pharmacology and diagnostic performance, secondly because – as stated above – pharmacody-

Table 8.1. The classic four phases for clinical testing of a new drug

Phase	Number of subjects	Type of subjects	Aims
Phase 1: clinical pharmacology and toxicity (initial administration to humans)	Tens	Healthy volunteers or patients	To obtain initial information on safety and tolerability (side effects) and on pharmacokinetics* on a wide range of doses. If phase 1 is conducted on patients, preliminary information on pharmacodynamics** can also be obtained. These are fundamental data for designing a phase 2 study
Phase 2: initial clinical study (first controlled studies)	Tens-hundreds	Patients	*Phase 2a, proof of concept:* to demonstrate that the drug at high doses is active on important pharmacodynamic end-points, well established for small samples of patients. *Phase 2b:* to select the best dosage and administration regimen to be used in phase 3. Possible secondary aims to obtain information on pharmacodynamics and therapeutic efficacy***
Phase 3: extended evaluation of treatment (further controlled and non-controlled studies)	Thousands (typically 2,000–5,000 for each arm of randomization)	Patients	To confirm safety and pharmacodynamic action as well as to demonstrate therapeutic efficacy on a sample representative of patient population, preferably using clinical end-points***. The study design implies randomization to a group of patients treated with the experimental drug and to a control group (patients treated with the standard therapy or a placebo)
Phase 4: surveillance after the drug enters the market	Thousands and over	Patients	To confirm safety, pharmacodynamic action, and therapeutic efficacy after drug registration for approved indications, frequently with comparison with other treatments. Studies on economic impact. Drug surveillance

* *Pharmacokinetics* quantitatively studies drug uptake, distribution, metabolism, and clearance, i.e. the effects of the action of the human body on the drug.
** *Pharmacodynamics* studies the biochemical and physiological effects of the action of the drug on the human body.
*** In phase 2 and, more frequently, in phase 3, special studies can be conducted on elderly patients, different ethnic groups, patients with kidney or liver disease, or on the interaction with other drugs, food and/or water [Bacchieri and Della Cioppa, 2004].

A more active role by the radiologists

namic end-points are increasingly based on imaging findings. Here radiologists should play a more active role, and not simply suppliers of data managed by other professionals to demonstrate drug activity and efficacy.

8.2. Study Classification

Studies[1] can be firstly classified as observational or experimental. While observational studies can be prospective or retrospective and longitudinal or transversal (i.e. cross-sectional[2]), experimental studies are necessarily prospective and longitudinal (Table 8.2).

Experimental or observational studies

Prospective or retrospective studies

Here we need several definitions [Altman, 1991]. Studies which evaluate one or multiple groups of subjects without any modification of the context of the events by the test driver are defined *observational*. When the measured events happen after subjects have been enrolled, the study is *prospective*. When the measured events happened before the subjects have been enrolled, the study is *retrospective*. In other words, in prospective studies enrollment precedes the measured events, in retrospective studies enrollment follows the measured events.

IRB approval and informed consent

Epidemiologic studies of disease incidence are typically observational. Distinguishing between prospective and retrospective observational studies is also crucial for regulatory issues of radiologic studies. For both study types, approval by the Ethics Committee (or Institutional Review Board, IRB) is an absolute prerequisite. Moreover, for prospective studies, informed consent to participate in the study must be obtained by each enrolled subject. For retrospective studies, IRB approval dispenses with the need for informed consent, under the obvious condition that data and images reported in possible publications allow the subjects to remain anonymous (see Chapter 10).

Prospective versus retrospective design

To understand the difference between prospective and retrospective studies, consider the following example. Suppose you want to estimate the prevalence of arterial hypertension in the patients admitted to a hospital. You could enroll all the patients admitted to the hospital during a defined time interval and measure their arterial pressure at the first visit (*prospective study*). Otherwise,

Table 8.2. General classification of studies

Observational				Experimental
Retrospective		Prospective		Prospective
Longitudinal (case-control)	Transversal (cross-sectional)	Longitudinal (cohort)	Transversal (cross-sectional)	Longitudinal randomized

[1] Here we refer to studies on humans but many concepts can also be applied to cell or animal testing, i.e. on biological systems evolving in time. The design of studies on phantoms is simpler, generally observational and transversal.

[2] The use of the term *cross-sectional* for transversal studies is ambiguous: in radiology this term is used for imaging modalities which provide images of body slices, i.e. *tomographic* techniques.

you could obtain the same information using the first arterial pressure reported in the medical record of all patients admitted to the hospital during a previous time interval (*retrospective study*).

Similarly, suppose you want to estimate the prevalence of the azygos lobe in subjects who undergo chest x-ray examination. You could enroll all the subjects with indications for a chest radiograph during a defined time interval and record all those who have this anatomic anomaly (*prospective study*). Otherwise, you could re-evaluate all the digital images stored in the picture archiving and communication system (PACS) obtained for the subjects who have undergone a chest x-ray examination in a previous time interval (*retrospective study*).

Radiologic retrospective studies

Retrospective radiologic studies require particular specification. There is a relevant difference between the use of the reports produced when the examinations were originally performed and the re-evaluation of images previously obtained (this second option offers the possibility of multiple readings and of measuring intra- and interobserver reproducibility – see Chapter 7). Note that a study evaluating the original readings (the previous reports) remains retrospective, even though it uses the reports prospectively produced when the examination was performed: it refers to events (the examinations) which happened before the decision to do the study. However, the prospective routine reading could be less meticulous (or the routine reports might have omitted non relevant findings, even though they were detected) than a new *ad hoc* reading performed by a motivated radiologist. We can test this hypothesis comparing the prospective reading (the original reports) with the re-evaluation. In any case, all these studies are retrospective.

Longitudinal versus transversal studies

Studies aimed at investigating the variation of a variable over time are defined *longitudinal.* At least two measurements are needed for each subject, commonly in temporal relation with an event (for instance, the administration of a drug or surgical treatment) which subdivides the time context into *before and after*. When the variable is measured only once for each subject (typically in opinion polls), the studies are defined *transversal*. This term highlights that the measurement takes a picture of the situation at a single point in time.

Finally, let us reiterate the concept that while observational studies can be prospective or retrospective and longitudinal or transversal, experimental studies are necessarily prospective and longitudinal.

8.3. Experimental Studies and Control Group

Experimental design

When the conditions in which the events happen are modified by the test driver according to a planned scheme, a study is defined *experimental*.[3] In general the purpose is to obtain information on the action or the efficacy of a treatment by making it emerge from the background, i.e. from the variability present in the population under investigation. A sample of subjects who are administered

[3] Note that the adjective "experimental" is used here to highlight a basic feature of the study design, while in a different context it is used either to define the research hypothesis we would like to "demonstrate", the H_1 hypothesis, opposite to the H_0 hypothesis or null hypothesis (see Chapter 3), or to define cell, animal, or phantom testing (see Chapter 10).

an experimental treatment (commonly a new therapy) is compared with a sample of subjects who are administered the standard treatment or a treatment simulator (placebo). This second sample of subjects is named *control group*. When patients are assigned to the experimental or control group by means of a randomization procedure, the study is defined *randomized controlled trial*.

Control group

Why can we not simply administer the new treatment to a group of patients and observe the events like we typically do in a phase 2 study? The answer is that *an improvement in one group of patients alone does not allow us to draw reliable conclusions in favor of the efficacy of a treatment.*

Regression to the mean

A well known phenomenon explaining this statement is the *regression to the mean*. Many diseases cause symptoms which change over time, without a steadily progressive course, even when the disease is not treated. In practice, patients exhibit relapsing and remitting phases or, at least, phases with heavy symptoms and phases with light symptoms. A key point is that the probability of the patient asking for a medical diagnosis or treatment is higher in the relapsing phase. If a remitting phase follows, we have a regression to the mean of the clinical status. Even in the absence of treatment, patients in this phase show an improvement, even though they will show a worsening in the future. In this experimental scheme (*before/after* on a single group of patients), even completely ineffective treatment could appear as effective. This phenomenon is highly evident for disease with a seasonal trend (e.g. peptic ulcer, allergic asthma). A control group allows us to see whether patients undergoing standard treatment or placebo have a clinical course similar to that observed in patients undergoing experimental treatment. Only this comparison enables us to demonstrate that the experimental treatment is more effective than the standard treatment or placebo. Obviously, we might also see the opposite result, i.e. the experimental treatment is less effective than the standard treatment or placebo.

Randomization

Once this standpoint has been accepted, the basic problem is how to assign each patient to the experimental or control group. This assignment must be *randomized*. Neither the patient nor any of her/his relatives, the physician or any of the other members of the medical team should play any role in determining the attribution to either the experimental or control group.

What happens without randomization?

This hard rule is extremely important and opens serious problems, also for ethical concerns. When a randomized controlled trial is done for a new anticancer treatment, the new (potentially more effective) treatment is withheld from all the patients randomized to the control group. On the other hand, if the patients were assigned in an *open modality* (i.e. if patients and/or physicians could freely choose the assignment), the experimental group would probably be larger and, more importantly, composed of patients with more advanced stages of disease than the control group. The paradoxical result would be that the patients undergoing the new treatment would show a worse clinical course than those of the control group, which would thus lose its control function. We would conclude in favor of the standard treatment or of the placebo, making a major (false negative) error. In such circumstances an effective drug could be judged ineffective.

A negative example (the Canadian screening)

Events such as this can actually occur. Some years ago, the negative results of a Canadian breast cancer screening program provoked widespread debate [Andersson et al, 1988]. The findings were not in favor of screening mammography and some authors even hypothesized that mammographic breast

compression could have worsened the clinical course in the women with breast cancer. Upon critical analysis [Baines et al, 1990; Di Maggio, 1992] many technical and methodologic flaws were found in this screening program. One of the several problems was that a number of symptomatic women (with a palpable lump) who had asked spontaneously for a mammographic examination appeared to have been assigned by nurses to the experimental group (women who underwent mammography). In a case like this, the good intentions of the nurses to accelerate the examination in symptomatic women doomed the experimental design to an unavoidable failure: a higher number of cancers were found in the screened women and these tumors were more advanced than those in the non screened women (the control group). What should the nurses have done? They should have admitted the symptomatic patients to an immediate clinical mammography (complete with physical examination, ultrasound, and needle biopsy, if needed), without placing them in the experimental screening group. Beyond the specific controversies which were aroused [Baines et al, 1990; Miller et al, 1991; Baines, 1994; Tarone, 1995; Bailar and MacMahon, 1997], this episode clearly demonstrates the negative effects which can arise from flaws in randomization.

Unbiased randomization is crucial. Only in this way can we minimize possible bias and make the experimental group and the control group as similar as possible. Only in this way can the administration of the new drug in the experimental group and the standard treatment or placebo in the control group become the main source of difference between the two groups.

A well-known problem: the signal-to-noise ratio

Douglas G. Altman [Altman, 1991] shrewdly suggests that this issue is similar to that of signal-to-noise ratio, well known in radiology as a main factor determining image quality. *Biological variability is the background noise on which we try to distinguish the signal, i.e. the effect of the experimental treatment.* If variability is high and the experimental treatment is not "miraculous", the only possibility available to us is to make the background noise as homogenous as possible between the two groups. In this way, the signal (the effect of the experimental treatment) will appear as the only main difference between the two groups. Every inhomogeneity between the two groups acts as a confounding factor, reducing the possibility of detecting the signal.

Inhomogeneity between two groups as a confounding factor

In radiology, when a contrast agent is intravenously (or also intra-arteriously) administered to study a vascularized lesion or a vessel (as in x-ray angiography, CT angiography, or MR imaging and angiography), radiologists use a technical procedure to make the signal-to-noise ratio higher – *image digital subtraction*. This procedure is effective only if the precontrast images (which work as a mask) are entirely equal to the postcontrast images with the single exception of the contrast-enhanced lesions and the vessels, which will appear bright on a dark background (or vice versa) on the subtracted images. If other differences are present between the precontrast and postcontrast images, the subtracted images are burdened by artifacts (if the patient moves, we have an artifact produced by insufficient image coregistration).

Intraindividual comparison

High homogeneity between experimental group and control group can be obtained by comparing the two treatments in the same subjects, when each patient is the control of herself/himself. However, this intraindividual design is possible only when the disease can be treated in a different way in the same subject. Examples include topical dermatologic or ophthalmologic therapies

(right arm treated with drug A, left arm treated with drug B; right eye treated with drug A, left eye treated with drug B) or the study of a chronic disease where drug A can be tested in a first time period and drug B in a second period in the same patient, with a washout period between treatments (cross-over studies). Later on, we will see that an intraindividual design can be used for prospective radiologic studies on diagnostic performance.

Historical control groups

Lastly, note that a control group is also very useful for observational studies, even though choosing *historical control groups* (data retrospectively obtained from series of patients not treated with the experimental treatment) is a very difficult task [ALTMAN, 1991].

8.4. Observational Studies

When the experimental design is not possible

There are situations which do not allow experimental studies to be carried out. We cannot prospectively randomize healthy subjects to the exposure or non exposure to a harmful substance in order to demonstrate that in exposed subjects we observe a higher incidence of the substance-related disease. Here, the only methods for demonstrating the association between the harmful substance and the disease (and to infer a causal relationship) are *observational studies, typically used in epidemiology*.

Cohort or follow-up studies

A first method is the *prospective longitudinal observational study*, also called *cohort study* or *follow-up study*. The efficiency of this study depends on the frequency of the expected events. Inclusion and exclusion criteria for patient enrolment are fundamental. Serious problems can be created by the temporal extent of the study: a great number of subjects can be lost at follow-up or initial conditions which favored the enrollment may significantly change. Moreover,

Surveillance bias

groups of subjects with different risk of disease could undergo different surveillance protocols: subjects at higher risk would have a more intensive surveillance with a higher probability of an earlier disease diagnosis (*surveillance bias*).

Case-control studies

A second method is the *retrospective longitudinal observational study*, also called *case-control study.* We identify a group of patients affected with the disease and a group of subjects not affected with the disease. From the history of each subject in both groups we try to evaluate whether one or multiple factors contributed to disease pathogenesis in the *cases* or prevented the development of disease in the *controls*.

A basic aspect of case-control studies is the selection of a suitable control group (which should be very similar to the group of cases except for the presence of the disease). This goal is sometimes reached by *matching* each case

Matching

with a control very similar to the case (for example for age and sex), obtaining couples of *paired* subjects. However, this method prevents us from investigating the role of these variables (age and sex) in disease pathogenesis.

Studies on NSF

A current example is given by the reports on *nephrogenic systemic fibrosis* (NSF), a disease associated with the intravenous administration of Gd-based paramagnetic contrast materials in patients with chronic kidney disease (CKD) in stage III or, more frequently, stage IV or stage V [TAMBURRINI ET AL, 2007]. Studies describing only NSF patients are undoubtedly less useful than those [SADOWSKI ET AL, 2007; RYDAHL ET AL, 2008, PRINCE ET AL, 2008] which describe both NSF patients (the cases) and CKD patients who received Gd-based contrast

material and did not develop NSF (the controls). Only this comparison can provide information on cofactors in disease pathogenesis.

Other critical aspects of case-control studies are as follows:

– selection of cases;
– better possibility of investigating the history of cases than that of controls (*recall bias*: cases remember the exposure to risk factors, controls do not remember them);
– general lower accuracy of retrospective investigation;
– different intensity of surveillance protocols (*surveillance bias*), as stated above for prospective longitudinal observational studies.

Recall bias

Surveillance bias

A third method is offered by *transversal* (*cross-sectional*) *observational studies*. Here we do not have any comparison between cases and controls. Information is obtained only once and does not regard longitudinal history. When we investigate actual events (e.g. presence or absence of a life habit, as with a *survey*) or events immediately following enrollment (e.g. the result of a diagnostic examination), the study design is *prospective*. When we investigate single past events without case-control comparison, the study design is *retrospective*. Critical aspects of transversal observational studies are as follows:

Transversal studies

Surveys

– the sample selection (if the subjects are volunteers, we can have *volunteer bias*);
– the response rates;
– the evaluation of cause-effect relations between the variables under investigation.

Volunteer bias

As a general warning, bear in mind that observational studies enable us to detect possible associations between events. When ethically possible, a more reliable evaluation of the incidence of these events and, above all, the inference of cause-effect relations should be done by means of experimental studies, i.e. prospective, longitudinal, randomized studies. Among the observational studies, cohort studies (prospective longitudinal) are commonly more reliable and less biased than transversal and case-control studies.

Randomized studies as a first choice

8.5. Randomized Controlled Studies: Alternative Approaches

In the previous sections, we referred to randomized controlled trials as designed using *parallel groups*. Alternative approaches can be used. Here we list the most important ones [Altman, 1991]:

Types of randomized controlled trials

1. *cross-over*, when all the patients receive both treatments, one before the other, with a randomized order of priority. Limitations are as follows: enrolled subjects can withdraw (*drop-out*) after the first treatment (e.g. because of side effects); when the effect of the first treatment can still be present after the administration of the second treatment (*carry-over effect*), we need to introduce a *washout time interval* between the evaluation of the effect of the first treatment and the administration of the second treatment; the cross over design can only be used for chronic diseases and when the effect of the treatment we are investigating is relatively rapid;

Cross-over design

Intraindividual paired data

2. *intraindividual paired data*, when each subject receives both treatments at the same time (only for topical therapies of double organs or skin areas – see Section 8.3). This scheme is frequently used for comparative radiologic studies;

Matched paired data

3. *matched paired data*, when pairs of subjects are set up for defined factors (e.g. age, sex, or other prognostic factors);

Sequential design

4. *sequential*, when a study on parallel groups is carried out until one of the two treatments proves to be significantly better than the other treatment (the overall results are calculated and evaluated after every new patient enters the study);

Factorial design

5. *factorial*, when all the possible combinations among treatments are evaluated with different groups of patients. For instance, for three treatments (A, B and C), we have six parallel groups of patients, treated as follows: A alone, B alone, C alone, A+B, A+C, and B+C.

8.6. Studies on Diagnostic Performance: Classification

Studies on diagnostic performance

Studying diagnostic performance always implies at least one comparison between the results of an examination and a *reference standard* (typically histopathology) which supplies the *truth* for defining whether a positive or negative result of the examination is true or false. Not in all studies, nor in all patients enrolled in a study, is the reference standard given by histopathology. It may be another diagnostic examination (e.g. the examination considered the *standard of care* up until the time of the study design) or a combination of histopathology for the positive cases and clinical and/or radiologic follow-up for the negative cases.

To evaluate diagnostic performance, two general variants of study design can be adopted: *non-comparative studies* or *comparative studies*. Moreover, comparative studies can be inter- or intra-individual, as shown in Table 8.3.

Non-comparative studies

Non-comparative studies are apparently simple: each result of the examination is compared with the reference standard. Hence, *a comparison really exists also in non-comparative studies*, but only with the reference standard. Obviously, the reading of examinations needs to be performed independently from the reading of the reference standard and vice versa. *This blinded reading is not at all a common event in clinical practice.* However, in radiologic prospective studies, the blindness can be favored by the physical separation between the radiology department and the pathology department.

Comparative studies

Comparative studies – increasingly performed due to the multiple diagnostic options offered by the technologic evolution of medical imaging – are more complex. When two (or multiple) imaging modalities are compared with each other, but each modality is performed in a different group of patients, we have an *interindividual comparative study*. When two (or multiple) imaging modalities are compared with each other but each modality is performed in the same group of patients, we have an *intraindividual comparative study*. In both settings, three conditions need to be met:

1. *randomization* of patients to different imaging modalities (interindividual studies) or of the temporal sequence of the different imaging in each patient (intraindividual studies);

Table 8.3. Studies on diagnostic performance

Study type	Description	Measures	Example: accuracy in diagnosing liver metastases
Non-comparative	Examination A versus RS in a series of patients	Diagnostic performance of A	CT versus IOUS
Comparative	In different groups (interindividual design): patients randomized to group I (A versus RS) or to group II (B versus RS)	Diagnostic performance of A Diagnostic performance of B Comparison between the diagnostic performance of A and B	CT versus IOUS MR versus IOUS CT versus MR
	In the same patients (intraindividual design): examinations A and B in each patient versus RS; randomization of the temporal sequence (group I: A before B; group II: B before A)	Diagnostic performance of A+B Comparison between the diagnostic performance of A and A+B Comparison between the diagnostic performance of B and A+B	CT+MR versus IOUS CT versus CT+MR MR versus CT+MR

RS = reference standard; CT = computed tomography; IOUS = intraoperative ultrasound; MR = magnetic resonance.

2. *independent reading of each examination from that of the reference standard*;
3. *reading of each examination independent from the reading of the other examination(s)*.

The first condition explains how *randomization* is a pivotal factor for all prospective comparative studies and how non-randomization in all retrospective comparative studies is a substantial limitation, being a potential source of relevant bias. The second condition is the same needed for non-comparative studies. The third condition, *the independent reading among the examinations*, implies that different radiologists interpret the examinations, each of them blinded to the results obtained by her/his colleagues. When the technical implementation of the examinations is standardized and the number of patients is high enough for a single reader not to recognize the cases at a repeated evaluation, an independent reading of one or multiple examinations of the same patients can be obtained by a single radiologist. However, two conditions need to be met:

1. a *mental washout period* lasting at least one or two weeks is required to prevent the reading radiologists from remembering the cases already evaluated. If during the washout period the radiologist reads other cases not included in the study, the efficiency of the mental washout is increased;
2. the examinations need to be presented in *random order*.

Single, double, triple blindness

Study blindness has some special features in radiology. The classic definition of single, double, or triple blindness is quite simple for randomized controlled trials. We have *single blindness* when only the patient [Altman, 1991] or only

the physician [Motulski, 1995] does not know who has been assigned to the experimental or control group. We have *double blindness* when both the patient and the physician do not know who has been assigned to the experimental or control group. When also the reader responsible for the evaluation of the effect of the diagnostic examination does not know who has been assigned to the experimental or control group, we have *triple blindness*. In the latter case, we have a differentiation between the physician administering the treatments and the physician evaluating the clinical status of the patients.

Reading blinded to demographics, clinical history, and previous examinations

For studies on diagnostic performance, blindness of reading radiologists can be required not only with respect to the reference standard but frequently also with respect to demographics, clinical history, and previous imaging or laboratory examinations. The latter requirement is aimed at determining the individual contribution of the imaging modalities under investigation but has the intrinsic limitation of being greatly different from the usual reading in clinical practice. In fact, in clinical practice each examination is positioned within a diagnostic sequential flow-chart for the planning of multiple examinations. Moreover, exact knowledge of clinical history, results of previous examinations, and diagnostic queries is needed to perform many examinations typically required as further workup.

Greater power of the intraindividual design

The comparison of the diagnostic performance of multiple examinations is generally more powerful when using an intraindividual rather than an interindividual design. In fact, when performing multiple examinations in the same patients, we obtain a reduction in variability which could be obtained only with a much larger number of patients using an interindividual design. The greater power of the intraindividual design compared to the interindividual design makes possible not only a reduction in the number of subjects, but also in financial costs and duration of the study. Statistical analysis takes into consideration this important difference: tests for paired data should be used for the intraindividual data (e.g. the McNemar test); tests for independent data should be used for the interindividual data (e.g. χ^2 and Fisher exact test). See Chapters 4 and 5 on this matter.

Note that comparative studies can evaluate the performance of more than two imaging modalities or techniques. If we compare three modalities, we need three groups of patients for an interindividual study and the performance of all the three examinations (in randomized order) in all the patients for an intraindividual study.

How can we combine this classification of the studies on diagnostic performance with the general scheme we presented in Table 8.2? The answer is quite complex.

Randomization as a marker of experimental studies

In our opinion, from a scientific viewpoint, all studies using some form of randomization should be considered as experimental studies, including intraindividual radiologic studies with randomization of the sequence of the examinations or with randomization of the reading order. In fact, the use of a randomization scheme implies that the test driver has introduced an experimental modification to reduce some variability or some source of bias as well as to increase the probability of observing a difference in diagnostic performance.

A possible classification of radiologic studies can be summarized as follows:

- retrospective studies are considered observational;
- longitudinal prospective cohort studies (e.g. following a cohort of subjects who undergo periodic screening examinations, without a parallel control group) are considered observational;

– transversal comparative intraindividual prospective studies with randomization of the sequential order of the exams are considered observational;
– controlled longitudinal studies with randomization to an experimental group or to a control group (e.g. periodic screening examinations versus no screening) are considered experimental;
– transversal comparative interindividual studies with randomization to an experimental group of subjects who undergo an imaging protocol and a control group who undergo another imaging protocol (e.g. a new technique versus a standard technique; CT versus ultrasound; MR versus CT etc.) are considered experimental.

Radiologic studies measuring impact levels higher than that of diagnostic performance

Lastly, we should remember that studies on diagnostic performance are only the second of the six ranks of efficacy of radiologic studies reported in Table 0.1. At high levels of the scale, especially at the fifth (impact on *outcome*) and at the sixth (societal impact), a diagnostic examination has to be considered as a *treatment* of which we are testing the efficacy with standard methods of clinical research.

8.7. Randomization and Minimization

"Random does not mean the same as haphazard"

We have already emphasized that randomization is a crucial feature of most studies. Douglas G. Altman [Altman, 1991] highlighted that "*random does not mean the same as haphazard*". To randomize means to assign a subject to a treatment (or an imaging modality, or an order of reading) according to a defined probability, which is usually the same for the different groups (0.5 or 50% in the typical case of two groups), without any possibility of predicting the assignment of each subject.

In contrast to what common sense would suggest, assigning patients in an alternate fashion to two imaging modalities (e.g. patients 1, 3, 5, 7 etc. to modality A and patients 2, 4, 6, 8 etc. to modality B) is not equivalent to randomization. The same can be said for the use of the ordinal numbers of the day, the week, or the month of enrollment as well as the date of birth of each patient.

Systematic allocation

All these types of *systematic allocation* of patients can be affected with bias and are therefore named *pseudo-random.* Another possibility of bias is using an open list of "random" numbers. One example is the telephone numbers on a page of the phone book of a large town. We match the first patient with the first number, the second patient with second number, the third patient with the third number, and so on. The patients with an even number undergo modality A, those with an odd number undergo modality B.

Pseudo-random allocation

What is the problem with these pseudo-random procedures? The problem is that we can predict the assigned group before enrolling each patient. This breaks a basic rule for correct randomization. The list of random numbers (as with the phone book) can be a valid method only if the list is blinded, that is to say if an independent person holds the list and tells the test driver the result of the matching between each patient and number only after their enrollment. Commonly, to ensure correct assignment in multicenter randomized trials, each center communicates each new enrollment to a central unit (e.g. a unit of epidemiology) which gives back the randomized assignment of the patient to one

of the groups. For single-center studies, we suggest the use of a computer program for generating random numbers to be consulted after each enrollment. In optimal conditions, the person consulting the computer should not be the test driver.

Why such high attention paid to randomization procedures?

Why should we pay such high attention to randomization procedures? The test driver may unconsciously alter the distribution of patients by typically placing patients with more severe disease or a higher suspicion of disease in the experimental group (which is given the new treatment or imaging modality). When the list of numbers is open, i.e. already known at the time of enrollment, the test driver can propose enrollment to a patient in a more or less convincing manner, guiding in such a way the patient's decision to give or not to give their consent to participate in the study according to the patient's condition and the known pseudo-random assignment.

Randomization may create inbalances among groups

However, even a correct randomization procedure can bring quantitative and qualitative distortions to the distribution between the groups. Let us consider the following example. Using a commercial software package we generate a sequence of 20 random numbers comprised in the interval [0, 9] to be used according to the following rule:

– the patients paired with a number from 0 to 4 undergo treatment A;
– the patients paired with a number from 5 to 9 undergo treatment B.

The sequence generated by the computer is:

4-2-7-8-3-5-0-9-1-0-9-2-5-5-6-7-8-4-9-7

We have 8 out of 20 numbers from 0 to 4 versus 12 out of 20 numbers from 5 to 9 (note that the odd/even distribution is also imbalanced: 11 odd versus 9 even). Therefore, a *simple randomization* procedure can produce an imbalanced distribution, especially when the sample of patients is small. The imbalance can be not only quantitative but also qualitative, concerning patient characteristics. As a consequence, the study results could be biased. These imbalances can be corrected using special types of randomization.

Simple randomization

Block randomization

Block randomization (or *restricted randomization*) is performed according to a scheme which has a balanced distribution within each block. The number of blocks is usually a multiple of the number of the groups. A simple example is the use of 6 blocks, each of them composed of 4 patients, to randomize to treatment A or treatment B, as follows:

1. AABB
2. ABAB
3. ABBA
4. BAAB
5. BABA
6. BAAB

In practice, we take on a random number from 1 to 6 and we assign the first four patients according to the corresponding scheme. For the next four patients, we take on a second random number, and so on. At the end of the enrollment,

the quantitative imbalance is limited to a maximum of one or two patients. However, patient-by-patient, the assignment of the last one of each block can be predicted. A solution is not to reveal the block size to the test driver or to change the block size during the study.

Stratified randomization

To avoid distortions in the distribution of patient characteristics – e.g. age and sex, disease severity, comorbidities etc. – we can use *stratified randomization*. Lists of blocks are generated for each subgroup (namely each *stratum*). In a simple case, we use a block randomization scheme for the females and another block randomization scheme for the males. In more complex cases, when we want to control multiple factors, the number of subgroups is equal to the number of possible combinations of factors. If we want to stratify for sex and three age brackets, we need $2 \times 3 = 6$ subgroups. However, a large number of subgroups is not a good idea due to the small number of subjects who finally are enrolled in each subgroup.

Cluster randomization

A particular type of randomization is *cluster randomization*. Here we randomize not each subject but a group of subjects. These groups are commonly determined by the family, the town, the house district, the hospital etc. However, a preliminary analysis is needed to verify the absence of biases due to patient difference among the clusters. In some cases we can use *weighted randomization* to obtain a different sample size for each patient group.

Weighted randomization

Minimization

Finally, we can adopt a nonrandom approach which permits a balanced assignment considering multiple prognostic factors: this is known as *minimization*. It offers advantages in comparison with classic randomization, unless we have a very large sample size (i.e. hundreds of patient for each randomized group) which tends to give homogeneous patient groups.

How does minimization work? It tends to minimize imbalances at each new enrollment. A computer program decides what to do after having considered all the previous enrollments. Suppose you want to assign patients to treatment A or to treatment B taking into consideration sex (male versus female) and age (lower than 35 years versus equal to or higher than 35 years). Observe the following simulation:

1. patient 1: male, 30-y.o. Being the first, he can be indiscriminately assigned to treatment A or B. A random procedure decides to assign the patient to treatment A;
2. patient 2: male, 28-y.o. To balance the assignment of patient 1, the minimization procedure assigns patient 2 to treatment B;
3. patient 3: female, 50-y.o. Being the first female, the software uses a random procedure and assigns patient 3 to treatment B;
4. patient 4: female, 40-y.o. To balance the assignment of patient 3, the minimization procedure assigns patient 4 to treatment A;

and so on…

We recommend the use of a dedicated software package.

8.8. Sample Size

The importance of sample size calculation

The sample size calculation is one the most important aspects of a study. It should be done early during the phase of protocol definition. A correctly designed study should include the definition of statistical power and the calcu-

lation of the sample size already in the protocol submitted for approval to the Ethics Committee.

However, we have noted that most medical imaging studies, including those which are published in highly ranked radiologic journals, lack a calculation of the sample size (which is the main factor determining statistical power).

The authors almost always define the α error (mostly 0.05, i.e. 5%) in the final part of *materials and methods* section (subsection *statistical analysis*). As we stated in Chapter 3, this implies a 1:20 maximum probability of obtaining a false positive result, i.e. of reporting a real difference when it is only due a random effect of variability. Because most published studies presents one or more significant results, i.e. with $p < 0.05$, the problem of sample size and power does not come into question: the sample has provided significant differences which allows the null hypothesis H_0 to be rejected and the experimental hypothesis H_1 to be accepted. *If a result is positive, it can only be true positive or false positive*. And the probability of a false positive result is just the obtained *p* value. *If the observed p value is very small* (say lower than 0.01), *the sample size and power might have been excessive, thus wasting time and money.*

When *p* is less than 0.05

When a study reports no significant differences, the problem of sample size and power comes into question. In fact, *without a preliminary calculation of sample size and power, we should take into consideration the possibility that the number of enrolled subjects is too small to demonstrate a real difference as significant.* In this case, we can retrospectively calculate the study power. Because the β error should be between 0.2 (20%) and 0.1 (10%) and the power is equal to $1-\beta$, if the power proves to be much less than 0.8 (80%), we can state that *the lack of significance of the study is not conclusive*. The study should be repeated with a sufficient sample size and power. This important problem rarely appears in the literature because many studies with nonsignificant results are not published. The reasons for this *publication bias* are as follows:

When *p* is equal to or greater than 0.05

Pubblication bias

1. these studies are frequently not submitted for publication to the journals;
2. when they are submitted to the journals, they have a high probability of being rejected.

This is a self-reinforcing loop which increases the trend of the non publication of studies with nonsignificant results.

What is the role of sample size in determining the quality of a study? Suppose you compare the overall accuracy of two diagnostic examinations, A and B, for a given disease in two randomized different groups of patients, each of them composed of 20 patients. The accuracy of A is found to be 30% (6/20) while that of B is 50% (10/20). The difference in overall accuracy is 20 % (50%-30%) but the statistical analysis (χ^2 test) gives us $p = 0.1967$ (nonsignificant). We conclude that there is insufficient evidence in favor of the greater accuracy of B with respect to A. Then we study 100 patients with examination A and 100 patients with examination B, adding 80 patients to those already studied for each of the two groups. The overall accuracy is again 30% for A and 50% for B, generated now by a ratio of 30/100 for A and by a ratio of 50/100 for B. Again we have obtained a difference equal to 20% but the χ^2 test now gives $p = 0.0038$ (highly significant). We now conclude that there is strong evidence in favor of the greater accuracy of B with respect to A. In Table 8.4 you can see how by increasing the

The sample size influences the study quality

Table 8.4. Simulation of a series of interindividual comparative studies of the overall accuracy of two examinations for a given disease: examination A in a group of patients, examination B in another group of patients (after randomization)

Patients	Examination A		Examination B		*p**
	TP + TN	Accuracy	TP + TN	Accuracy	
20 + 20 = 40	6	0.30	10	0.50	0.1967
30 + 30 = 60	9	0.30	15	0.50	0.1138
40 + 40 = 80	12	0.30	20	0.50	0.0679
50 + 50 = 100	15	0.30	25	0.50	0.0412
60 + 60 = 120	18	0.30	30	0.50	0.0253
70 + 70 = 140	21	0.30	35	0.50	0.0157
80 + 80 =160	24	0.30	40	0.50	0.0098
100 + 100 = 200	30	0.30	50	0.50	0.0038

TP = true positives; TN = true negatives; * χ^2.

sample size, the *p* value progressively decreases (i.e. the significance increases), even though the overall accuracies of A and B are the same. Note that significance ($p < 0.05$) can already be obtained with a sample of 50 + 50 patients and a high significance ($p < 0.01$) with a sample of 80 + 80 patients. A sample size calculation would have defined an optimal size at 65 + 65 patients.

Preliminary definition of power and sample size

This example raises a provocative question. *If we progressively increase the sample size, will we reach statistical significance for an existing difference? Even if the difference is very small?* The simple answer is: *Yes, but we should not do so. The golden rule for a good research study is to define a priori sample size and power.* Three facts testify in favor of this rule:

1. in large clinical randomized controlled trials sample size and power are always calculated a priori;
2. some highly ranked journals publish the research project and protocol of a clinical trial as a self standing article, submitted before starting enrollment;
3. during a randomized controlled trial a partial data analysis can be performed (typically when half the sample has been enrolled), named *interim analysis*, but this should be planned a priori and, to be rigorous, an interim analysis should imply harder thresholds for significance (α error lower than 0.05) at the final analysis.

Interim analysis

Sample size calculation

How should the sample size be calculated? This task requires the cooperation of a professional statistician. However, bear in mind that a basic factor for calculating the sample size is the definition of the minimal difference thought to be clinically relevant which we want to demonstrate as statistically significant. This amount cannot be calculated with mathematical formulas. It should be derived from a critical analysis of the papers previously published on the matter under investigation and from an evaluation – necessarily subjective – of the clinical and scientific context. This can only be done by the radiologist(s) conducting the study.

Standardized difference

Sample size calculation is technically based on the *standardized difference*, equal to the ratio between the minimal difference thought to be clinically rele-

vant (δ) and the standard deviation (s) which quantifies sample variability. Therefore, standardized difference = δ/s.

The larger the standardized difference, the smaller the sample size, and vice versa. In fact, with a given standard deviation, the larger δ (numerator) is, the higher the probability of obtaining a significant difference with a lower number of patients. Similarly, with a given δ, the lower the standard deviation is, the lower the overlap between the effects of the two compared treatments and, again, the higher the probability of obtaining a significant difference with a lower number of patients.

For the comparison between continuous variables in two independent groups, we have to define not only δ/s but also α error and power (1-β; usually from 0.8 and 0.9). The use of a nomogram such as proposed by Douglas G. Altman [ALTMAN, 1980] and shown in Figure 8.1 makes possible a rough calculation having at hand the above mentioned parameters.

For the comparison between continuous variables for paired data, we should consider not the standard deviation of the measured data but the standard deviation of the differences between the measured data for each pair of data in the same subject (s). Here the standardized difference is equal to 2δ/s. Again, Altman's nomogram shown in Figure 8.1 makes possible a rough sample calculation.

For categorical data, if we define the expected proportion of events in the two samples as p_1 and p_2, the standardized difference is equal to:

$$\frac{p_1 - p_2}{\sqrt{p_m(1-p_m)}}$$

where: $p_m = (p_1 - p_2)/2$

Using this definition, we can use Altman's nomogram shown in Figure 8.1 for categorical data as well.

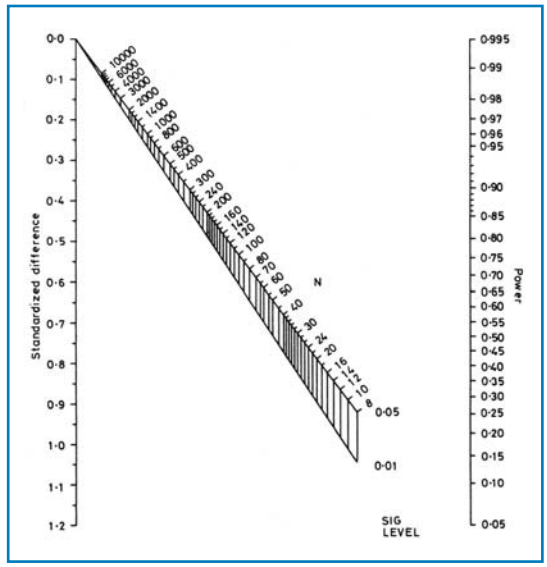


Figure 8.1. Altman's nomogram for calculating sample size or statistical power. Drawing a straight line from the value of the *standardized difference* to that of the *power*, we can calculate the *sample size*. Drawing a straight line from the value of the *standardized difference* to that of the *sample size*, we can calculate the *power*. From: Altman DG (1980) Statistics and ethics in medical research. III. How large a sample? Brit Med J 281:1336-1338 (with permission of the author and of the copyright owner [BMJ group]).

8.9. Systematic Reviews (Meta-analyses)

Multiple published studies on a certain matter make a mass of data and results available to the scientific community. Sometimes, the results may be inconclusive (e.g. insufficient statistical power due to small sample size) or conclusive but controversial. An example of the latter case is the presence of studies in favor of a superior diagnostic performance of A compared to B, studies considering A and B to be equivalent, and studies in favor of B compared to A. Note that:

a. the quality of the studies can be extremely different due to limitations in design and implementation;
b. the results of the studies are based on immediately available analytical data (from the Results section of the articles) or we need to ask the authors in person to provide their analytical data.

A study of a sample of studies

We may think of the already published studies on the subject we would like to investigate as a population of units to be potentially enrolled in a *study of a population of studies*. All the data contained in the studies judged of acceptable quality are a large dataset which can be analyzed with suitable statistical techniques (*meta-analysis*) to generate a new result, based on a larger sample than that included in the individual studies.

Systematic review

The authors of a *systematic review* should:

1. describe in detail the matter under investigation, i.e. define one (or more) clear end-point(s) (e.g. the diagnostic accuracy of examination A for disease X);
2. establish inclusion and exclusion criteria of the studies in the meta-analysis (*metaprotocol*);
3. utilize a *systematic* procedure for searching all the published studies on the matter under investigation (firstly, by means of websites available on the internet, using predefined key words; secondly, by means of the *References* of the studies initially found);
4. analyze the whole text of all the studies found and include in the meta-analysis only those corresponding to the defined quality criteria;
5. make new calculations on the new whole dataset to provide a *new result*;
6. conclude, if possible, giving a new and more precise estimate of the end-point (for the example of the accuracy of examination A for disease X, the confidence interval will be narrower than that obtained in the individual studies included in the meta-analysis, meaning a more precise estimation).

Metaprotocol

Using this method we can study the comparison between different treatments or between different imaging modalities or techniques. The mathematical and statistical techniques used for performing a meta-analysis, including the tests for homogeneity across the studies included, are specialized in nature and go beyond the limits of the current book.

Meta-analyses: pros

Meta-analysis has advantages but also intrinsic limitations. The following reasoning stands in favor of this approach. To further investigate controversial problems in clinical research, instead of a new big prospective study with high economic costs and a long time needed for obtaining the results (e.g. survival studies), we can *re-use* the data of already published high quality studies to

obtain new and more robust evidence. This advantage is maximal for the evaluation of the treatment of rare disease for which individual studies on very small samples cannot reach the power needed for a statistical demonstration of a therapeutic effect.

Meta-analyses: cons

Publication bias

The basic limitation of meta-analysis is related to the above mentioned *publication bias*. Because the publication of studies reporting statistical significances is more probable than that of studies reporting the absence of statistical significances, systematic reviews and meta-analyses reinforce this bias, presenting a *sum of results* emphasizing the bias in favor of positive results. However, the importance of systematic reviews in radiology is increasing. One of the most prominent journals in this field, *Radiology*, recently introduced a new series of articles entitled *Evidence-Based Practice* mainly composed of systematic reviews.

At any rate, we advise radiologists who want to enter this field as authors to work in close cooperation with professional statisticians.

Forest plot

A technical aspect we want to describe is the graphical representation of the results of a meta-analysis. This is very useful for grasping the high level of information and synthesis offered by this method. A good example is given by a recent meta-analysis on the diagnostic performance of breast MR imaging [Peters et al, 2008]. The authors identified a total of 1096 studies. Of them, 251 were eligible but only 44 were included (sample size, from 14 to 821 patients; breast cancer prevalence from 23% to 84%). From these studies the authors extracted data for 2,808 examinations in patients with breast cancer and 1,827 examinations in subjects not affected with breast cancer. The meta-analysis estimated the sensitivity of breast MR imaging equal to 0.90 (95%CI: [0.88, 0.92]) and the specificity equal to 0.72 (95%CI: [0.67, 0.77]). Figure 8.2 shows the *forest plot* of sensitivity and specificity of individual studies and of their overall estimate.

8.10. Levels of Evidence

Levels of evidence

The need to evaluate the relevance of the various studies in relation with the reported level of evidence has generated a hierarchy of the levels of evidence based on study type and design.

According to the *Centre for Evidence-Based Medicine* (Oxford, UK), studies on diagnostic performance can be ranked on a five-level scale, from 1 to 5 (Table 8.5). On the basis of similar scales, some authors distinguish four degrees of recommendations, from A to D (Table 8.6).

However, we should consider that today we have multiple different classifications of the levels of evidence and of degrees of recommendation. The same degree of recommendation can be represented in different systems using capital letters, Roman or Arabic numerals, etc, generating confusion and possible errors in clinical practice.

The GRADE system

A new approach to evidence classification has recently been proposed by the *GRADE working group* [Atkins et al, 2004] with special attention paid to the definition of standardized criteria for releasing and applying clinical guidelines. The GRADE system states the need of an explicit declaration of the

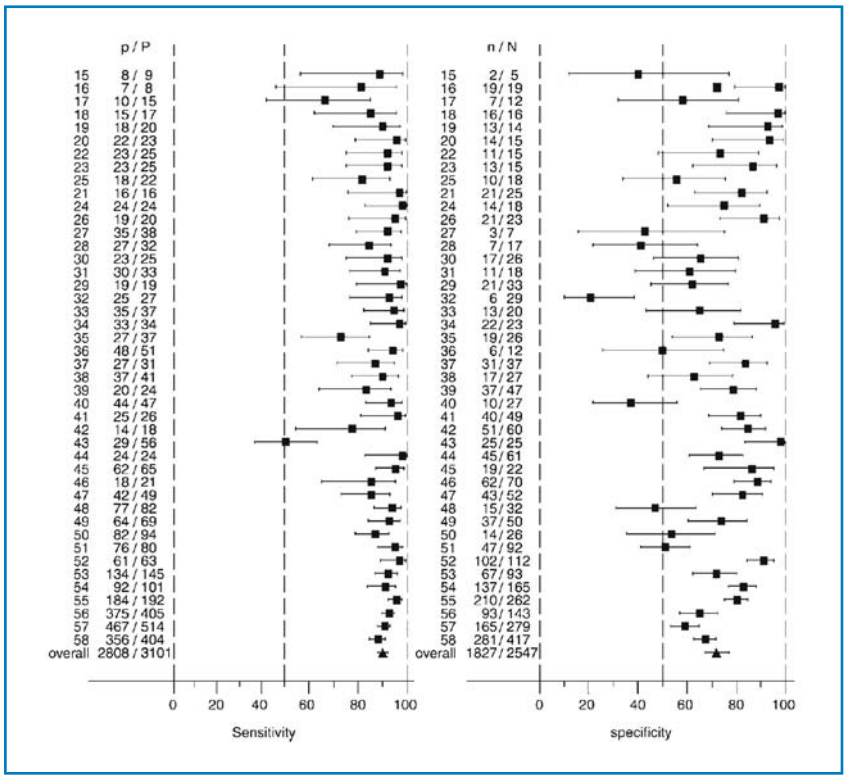


Figure 8.2. Graphical representation of the results of a meta-analysis. The authors systematically reviewed the studies on breast MR imaging in which sensitivity and specificity were evaluated. In both columns, the numbers on the left from 15 to 58 indicate the 44 included studies (as numbered in the references), the small black squares show the point estimate of sensitivity or specificity of each study, while the horizontal line crossing the small black square represents the 95% CI associated with this estimate; on the last line (overall) on the bottom of each of the two columns the new estimate and 95% CI of sensitivity and specificity is reported (with a triangle instead of a square). Note that the 95% CIs associated with the new estimates are reduced in comparison with most of the included studies. From: Peters NH, Borel Rinkes IH, Zuithoff NP et al (2008) Meta-analysis of MR imaging in the diagnosis of breast lesions. Radiology 246:116-124 (with permission of the author and of the copyright owner [RSNA]).

methodologic core of a guideline, with particular regard to: quality of evidence; relative importance, risk-benefit balance, and value of the incremental benefit for each outcome. This method, apparently complex, finally provides the following four simple levels of evidence:

– *high*, when further research studies are thought unlikely to modify the level of confidence of the estimated effect;
– *moderate*, when further research studies are thought likely to modify the level of confidence of the estimated effect and the estimate itself of the effect;
– *low*, when further research studies are thought very likely to modify the level of confidence of the estimated effect and the estimate itself of the effect;
– *very low*, when the estimate of the effect is highly uncertain.

Table 8.5. Levels of evidence of studies on diagnostic performance

Level of evidence	Study type
1a	Systematic reviews with homogeneous meta-analyses of level 1 studies Multicenter studies, in consecutive patients with a reliable and systematically applied reference standard, of diagnostic criteria previously established by explorative studies
1b	Single-center studies, in consecutive patients with a reliable and systematically applied reference standard, of diagnostic criteria previously established by explorative studies
1c	Studies of diagnostic examinations with very high sensitivity (*SNOUT*) and of diagnostic examinations with very high specificity (*SPIN*) *
2a	Systematic reviews with homogeneous meta-analyses of level 2 or higher studies
2b	Explorative studies of diagnostic criteria in cohorts of patients with a reliable and systematically applied reference standard; definition of diagnostic criteria on parts of cohorts or on databases
3a	Systematic reviews with homogeneous meta-analyses of level 3 or higher studies
3b	Studies of non consecutive patients and/or without systematic application of the reference standard
4	Case-control studies Studies with inadequate or non-independent reference standard
5	Experts' opinions without critical evaluation of the literature

Source: Centre for Evidence-Based Medicine, Oxford, UK (http://www.cebm.net/index.aspx?o=1025; accessed February 24, 2008), with modifications.
* For the definitions of SPIN and SNOUT, see Chapter 1.

Table 8.6. Degrees of recommendation

Degree of recommendation	Study type
A	Level 1 studies
B	Consistent level 2 or 3 studies or extrapolations* from level 1 studies
C	Consistent level 4 studies or extrapolations* from level 2 or 3 studies
D	Level 5 studies or low-quality or inconclusive studies of any level

Source: Centre for Evidence-Based Medicine, Oxford, UK (http://www.cebm.net/index.aspx?o=1025; accessed February 24, 2008), with modifications. **Extrapolation* is the translation of a study to clinical situations different from those of the original study.

Similarly, the risk-benefit ratio is classified as follows:

– *net benefit*, when the treatment clearly provides more benefits than risks;
– *moderate*, when the treatment provides important benefits, but there is a tradeoff in terms of risks;
– *uncertain*, when we do not know whether the treatment provides more benefits than risks;
– *lack of net benefit*, when the treatment clearly provides more risks than benefits.

The procedure gives only two levels of recommendations:

– *do it* or *don't do it*, when we think that the large majority of well informed people would make this decision;
– *probably do it* or *probably don't do it*, when we think that the majority of well informed people would make this decision but a substantial minority would have the opposite opinion.

As the reader can see, the GRADE system finally differentiates between *strong recommendations* and *weak recommendations*, making the application of the guidelines to clinical practice easier. Detailed information on this method can be found in the article by Atkins et al [Atkins et al, 2004].

Application of the GRADE system to diagnostic tests and strategy

Recently, the application of GRADE to diagnostic tests and strategies has been discussed [Schünemann et al, 2008], including a list of particular factors that decrease the quality of evidence for studies on diagnostic performance and how they differ from other interventions. When randomized controlled trials comparing the impact of alternative diagnostic strategies on "patient-important" outcomes are available, the general GRADE system can be directly used. When this is not the case, we should make inferences about the impact on patient-important outcomes from the available studies on diagnostic performance. In other words, *diagnostic performance is only a surrogate for patient-important outcomes*. To make inferences in favour of diagnostic tools implies the availability of effective treatment, reduction of test-related adverse effects or anxiety, or improvement of patient wellbeing from prognostic information. "The key questions are whether a reduction in false negatives (cases missed) or false positives and corresponding increases in true positives and true negatives will occur, how accurately similar or different patients are classified by the alternative testing strategies, and what outcomes occur in both patients labelled as cases and those labelled as not having disease" [Schünemann et al, 2008].

References

Altman DG (1980) Statistics and ethics in medical research. III. How large a sample? Brit Med J 281:1336-1338

Altman DG (1991) Practical statistics for medical research. London: Chapman & Hall 74-103

Andersson I, Aspegren K, Janzon L et al (1988) Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. BMJ 297:943-948

Atkins D, Best D, Briss PA et al for the GRADE working group (2004) Grading quality of evidence and strength of recommendations. BMJ 328:1490 (http://www.bmj.com/cgi/content/full/328/7454/1490. Accessed March 3, 2008)

Bacchieri A, Della Cioppa G (2004) Fondamenti di ricerca clinica. Milan. Springer 303-321

Bailar JC 3rd, MacMahon B (1997) Randomization in the Canadian National Breast Screening Study: a review for evidence of subversion. CMAJ 156:193-199

Baines CJ, Miller AB, Kopans DB et al (1990) Canadian National Breast Screening Study: assessment of technical quality by external review. AJR Am J Roentgenol 155:743-747

Baines CJ (1994) The Canadian National Breast Screening Study: a perspective on criticisms. Ann Intern Med 120:326-334

Di Maggio C (1992) The efficacy of mammography. Radiol Med 83:140-143

Filippi N, Grossman RI, Comi G (1999) Magnetic resonance techniques in clinical trials in multiple sclerosis. Milan. Springer

Hoffman JM, Gambhir SS, Kelloff GJ (2007) Regulatory and reimbursement challenges for molecular imaging. Radiology 245:645-660

Miller AB, Baines CJ, Turnbull C (1991) The role of the nurse-examiner in the National Breast Screening Study. Can J Public Health 82:162-167

Motulski H (1995) Intuitive biostatistics. New York, Oxford. Oxford University Press 184-185

Peters NH, Borel Rinkes IH, Zuithoff NP et al (2008) Meta-analysis of MR imaging in the diagnosis of breast lesions. Radiology 246:116-124

Prince MR, Zhang H, Morris M et al (2008) Incidence of nephrogenic systemic fibrosis at two large medical centers. Radiology 248:807-816

Rydahl C, Thomsen HS, Marckmann P (2008) High prevalence of nephrogenic systemic fibrosis in chronic renal failure patients exposed to gadodiamide, a gadolinium-containing magnetic resonance contrast agent. Invest Radiol 43:141-144

Sadowski EA, Bennett LK, Chan MR et al (2007) Nephrogenic systemic fibrosis: Risk factors and incidence estimation. Radiology 243:148-157

Schünemann HJ, Oxman AD, Brozek J et al for the GRADE Working Group (2008) Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 336:1106-1110

Tamburrini O, Balducci A, Anzalone N et al (2007) Fibrosi nefrogenica sistemica: raccomandazioni per l'uso degli agenti di contrasto a base di Gd. Documento SIRM-SIN-AINR (http://www.sirm.org/news/NSF_2007)

Tarone RE (1995) The excess of patients with advanced breast cancer in young women screened with mammography in the Canadian National Breast Screening Study. Cancer 75:997-1003

Therasse P, Arbuck S, Eisenhauer E et al (2000) New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 92:205-216

Therasse P, Eisenhauer EA, Verweij J (2006) RECIST revisited: a review of validation studies on tumour assessment. Eur J Cancer 42:1031-1039

WHO (1979) Handbook for Reporting Results of Cancer Treatment. World Health Organization Offset Publication No. 48. Geneva. WHO

9

Bias in Studies on Diagnostic Performance

He uses statistics as a drunken man uses a lamp post...
for support rather than illumination.

REX STOUT

Torture numbers,
and they will confess to anything.

GREGG EASTERBROOK

Bias or systematic distortion

Introducing Chapter 8, we emphasized that statistical testing of the null hypothesis H_0 and potential acceptance of the experimental hypothesis H_1 can be reliably done only if a study is not flawed by substantial *biases*, namely *systematic distortions* which could explain the observed difference as an alternative to a real difference between the compared samples.

Random or systematic?

The adjective *systematic* gives the term a particular meaning. Note that the presence of distortions in practical experiments cannot be entirely avoided. This is due to measurement errors or to other aspects of physical and biological variability. However, if these errors are randomly distributed, when the sample is large enough, they tend to annul themselves. They can generate background noise causing a more difficult detection of the signal, which here is the difference between two samples for a given characteristic. But this problem can be solved with a suitable study design and lastly with a preliminary calculation of the sample size. *If the distortion is not randomly distributed but systematic, we have a real bias which can determine a false result.*

Hard removal of a bias effect

The situation is similar to when the navigator of a car indicates the wrong location of the vehicle on the map on the videoscreen. The difference is that, in scientific research, we cannot look out the window to see where we really are. *The experimental data are the only messages we receive from reality*. If there is bias, the message is misleading. If the bias affects the study design, quantifying its magnitude in order to remove it from the data can be very difficult. In special cases, some statistical techniques can be used to try to remove the bias, but the result of these procedures is frequently questionable. If bias is due to a non-independent reading, all the

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

examinations can be reevaluated independently by new blinded readers. However, some bias can be unavoidable, often for ethical reasons. This must be recognized in the Discussion of the paper, in a subsection dedicated to *Study Limitations*.

In Chapter 8 we presented the study design and its variants, including systematic reviews, and the levels of evidence provided by the studies. There we gave you some idea regarding what to do in order to implement a good research study. In this chapter we present a list of errors, i.e. what not to do, namely the most important sources of bias in radiologic studies. Particular attention will be paid to the *studies on diagnostic performance*, level 2 of the hierarchical scale of the studies on efficacy of imaging modalities (see Table 0.1). However, the discussion is also widely valid for studies on *technical performance* (level 1), and *diagnostic impact* (level 3). For higher levels of efficacy (4, therapeutic impact; 5, outcome impact; 6, societal impact) we should also take into consideration other general aspects and statistical techniques of clinical research not included in the current book (e.g. measurement of clinical end-points, calculation of quality-adjusted life years, survival analysis, cost per saved life, etc.).

9.1. Classification

External and internal validity

We should distinguish between bias influencing the *external validity* of a study, that is the applicability of its results to clinical practice, and bias influencing the *internal validity* of a study, that is its inherent coherence. Bias influencing the external validity is due to *errors in planning the study*, those influencing the internal validity are due to errors in implementing the study (Figure 9.1). The

Planning and implementation

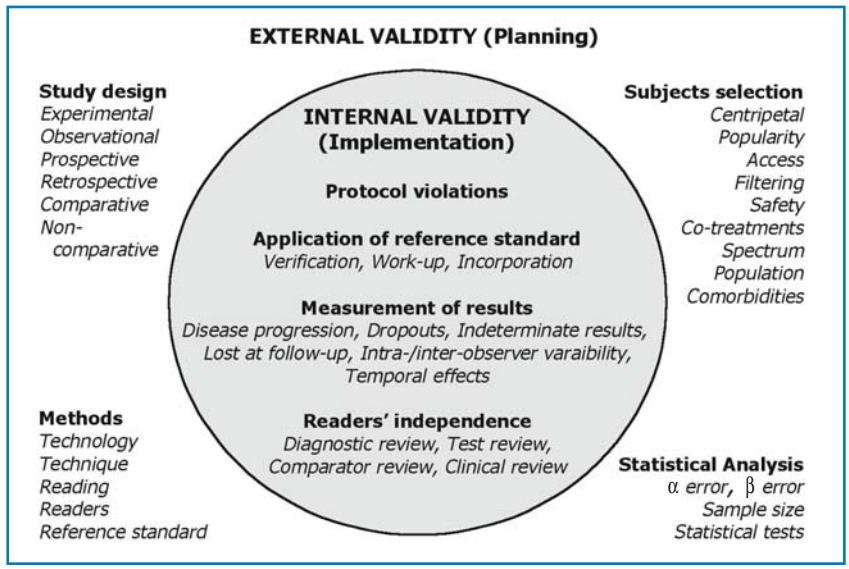


Figure 9.1. Synopsis of the sources of bias in studies on diagnostic performance. To apply the results of a study to clinical practice, it must have internal validity (no errors in implementation) and external validity (no errors in planning).

reader should pay attention to the distinction between external and internal validity. The two concepts are not independent of each other: *the internal validity is a necessary but not sufficient condition for a study to have external validity.*

All kinds of bias limit external validity

Thus, all kinds of bias influence the external validity of a study. However, while errors in planning have a negative effect on the external validity but possibly no effetc on internal validity, errors in implementation have a negative effect primarily on internal validity and secondarily on external validity. The lack of internal validity makes the results themselves unreliable. In this case the question about the external validity (i.e. the application of the results to clinical practice) makes no sense. As a consequence, only the results of a study not flawed by errors in planning and implementation can be applied to clinical practice [Kelly et al, 1997]. In fact:

- a study without bias from errors in planning but with important bias from errors in implementation produces intrinsically unreliable data, hence not applicable to clinical practice;
- a study without bias in implementation has no intrinsic contradictions but could be non applicable to clinical practice due to errors in planning;
- a study with bias from errors in planning and implementation not only provides intrinsically misleading results, but, also if repeated without errors in implementation, cannot be applied to clinical practice due to persisting errors in planning.

The reader will have realized that several items are present in both planning and implementation. Consider the reference standard: the error in planning is to choose an inadequate reference standard (*imperfect standard bias*); the error in implementation is an incorrect use of the planned reference standard. We can go the wrong way by either choosing the wrong rules or applying the right rules incorrectly (but also adding errors in the application of already incorrect rules). Indeed, there is probably only one right way to do a correct study but infinite ways to make errors that render the study worthless.

Infinite ways to make errors

A bias in implementation can be due to:

1. flaws in protocol application;
2. unforeseen events or events due to insufficient protocol specification;
3. methods defined in the study protocol which implied errors in implementation.

For items 2 and 3, the flaws in implementation depend in some way on errors in planning. This does not seem to be the case for item 1. However, if in a study we have many protocol violations, the study protocol was probably theoretically correct but only partially applicable. In other words, *bias in implementation frequently originates as an error in planning*.

9.2. Bias Affecting External Validity

This can be subdivided into four groups: (1) study design, (2) subject selection, (3) radiologic methods and reference standard, and (4) statistical analysis.

9.2.1. Study Design

See Chapter 8 for a detailed discussion about study design. Errors in study design determine commonly irrecoverable bias. From this point of view, always consider the substantial superiority (meaning lower probability of bias) of experimental versus observational studies as well as of prospective versus retrospective studies.

Potential bias of a retrospective comparative study

A simple example is the evaluation of a new technology in comparison with an old technology. Suppose we want to compare 64-row with 16-row multidetector CT (MDCT) in the diagnosis of hemodynamically significant coronary stenoses. In 2005 we installed a 16-row unit and performed 200 coronary studies with conventional coronary angiography (CCA) as a reference standard. In 2007 we installed a 64-row unit and performed 200 coronary studies, again using conventional coronary angiography as a reference standard. In both series, MDCT examinations were analyzed and reported before coronary angiography was performed. We have the impression that image quality generally increased and that the number of nondiagnostic MDCT exams decreased. Moreover, we remember many more cases of substantial correspondence between the MDCT report of the CCA findings. Using all these examinations, can a reliable retrospective study be done which compares 16-row MDCT and 64-row MDCT for the diagnosis of significant coronary stenoses?

The answer is: No, it is better to work on an alternative project. Why? There are many reasons:

1. We have no certainty that the two series of patients (those who received a 16-row MDCT and those who received a 64-row MDCT) are similar for presence of significant coronary artery disease (CAD) and for disease severity (number of vessels involved). For example, if more patients with high pretest probability of CAD are present in the second series, we have a potential bias in favor of 64-row MDCT, at least in terms of per patient sensitivity and positive predictive value;
2. Changes in the radiologic team working on MDCT before/after starting with the second series might have influenced the diagnostic performance creating a power *confounding factor* in favor of one of the two series;
3. In the first two years of experience with the 16-row equipment we have indubitably progressively increased our diagnostic performance according to an obvious learning curve, thus generating a bias in favor of the 64-row study;
4. During the four-year period of the retrospective study, the diagnostic performance of CCA might have changed (modifications of the medical team or of technology), determining a change in the reference standard between the two groups;
5. If CCA was performed knowing the MDCT results, there is a problem of incorporation of the MDCT result in that of CCA which serves as a reference standard, for both the two series. However, the situation would be even more marked if the CCA team started taking into consideration the MDCT results only after the installation of the 64-row unit (which would again be favored).

The reader will have realized that in these circumstances potential study defects can cause bias in patient selection, readers, reading, and reference standard. However, the bias in the study is not determined by our actions. It is instead the retrospective design which is burdened by a series of serious flaws,

similar to those we face when using an observational before/after design in a clinical phase 2 study with a new drug. Patient randomization to one of the two MDCT units is lacking. This reasoning holds also for other situations regarding the development of technology. Retrospective comparative studies which evaluate the performance of a new contrast agent in comparison with an old agent are burdened by the same defects and bias.

Possible solution to the problem

What should we do with our data on MDCT coronary angiography? We could write two separate articles, without a direct comparison between the new and the old technology. In the Discussion of the paper on the first series we would compare our results with those obtained by other authors with 16-row units, recognizing the technologic limitations of our study. In the Discussion of the paper on the second series with 64-row MDCT, we would comment that sensitivity and/or specificity appear better than those obtained using 16-row units (including our already published study) and compare our results with those obtained by other authors with 64-row units. In both papers we would acknowledge the limitation due to the retrospective design.

However, the best solution is to plan a new study on the diagnostic performance of the 64-row unit, preliminarily discussing all the aspects of the study design with a statistician, including the calculation of the sample size. *This is certainly the most promising hypothesis, where time and money can be more suitably invested.*

The method of a prospective comparative intraindividual (cross-over) study is not a valid alternative, especially for ethical reasons: this would mean the same patient would undergo both the 16-row and the 64-row MDCT with a double radiation exposure and a double contrast injection.

The last hypothesis is to randomize patients to the 16-row unit or the 64-row unit. However, another ethical concern renders this solution unfeasible. If the literature has already demonstrated that 64-row MDCT provides better diagnostic performance in comparison with historical data obtained with 16-row MDCT, the patients randomized to the old technology would have an unacceptably high probability of undergoing an examination with lower diagnostic performance. The IRB would not approve the study or patients might refuse to be enrolled. If our 64-row MDCT were one of the first installed in the world we could adopt a *sequential design* with planned data analysis after each enrollment to demonstrate the superiority of 64-row MDCT. Finally, we could design a phantom study (with relatively low economic costs) or animal testing (with much heavier economic costs).

As you can see, the study design is crucial for the scientific quality of a research study on diagnostic performance. This affirmation is even more valid for radiologic studies at higher levels of efficacy than those on diagnostic performance (see Table 0.1).

The limitations of retrospective studies are made clear by the following example.

Greater sensitivity does not imply a positive outcome impact

Contrast-enhanced breast MR imaging is well known as a technique characterized by a very high sensitivity and good specificity. This last feature was recently confirmed by a systematic review of 251 studies [Peters et al, 2008]. These authors selected 44 suitable studies and the meta-analysis provided an overall sensitivity of 0.90 (95%CI [0.88, 0.92]) and an overall specificity of 0.72 (95%CI [0.67, 0.77]). This diagnostic performance has a potential clinical impact for preoperative staging. In fact, breast MR imaging was demonstrated to be more sen-

sitive than mammography in detecting multiple malignant lesions (multifocal and multicentric cancers), a finding confirmed in the Italian multicenter study using the whole excised breast as pathologic reference [Sardanelli et al, 2004]. Moreover, breast MR imaging is also highly sensitive in detecting synchronous cancers at the contralateral breast [Lehman et al, 2007].

The clinical problem of preoperative breast MR imaging is now open. In fact, we should take into consideration that radiotherapy (and chemotherapy, when performed) have a high probability of curing the malignant lesions undetected with conventional imaging. Thus, whether preoperative breast MR imaging has a positive impact (less recurrences and less contralateral cancer, higher quality of life, higher survival), a null impact, or a negative impact (overdiagnosis and overtreatment, more mastectomies, more aggressive surgery) in breast care is a matter of debate.

In this context, in 2004 Fischer et al. published an extremely interesting article [Fischer et al, 2004]. They compared 121 patients who performed preoperative breast MR imaging (MR group) with 225 patients who did not undergo preoperative breast MR imaging (non-MR group). Both groups were followed up for a mean time of about 41 months. The rate of conservative surgery was 71.1% for the MR group and 61.3% for the non-MR group; the rate of ipsilateral recurrence 1.2% and 6.8%, respectively; the rate of contralateral tumors was 1.7% and 4.0%, respectively. For the latter two comparisons, the statistical significance was very high ($p < 0.001$). The question is: Are the results of this study (almost 10% more conservative surgery and a highly significant reduction in local recurrence and contralateral cancers in the MR-group) clearly conclusive in favor of preoperative breast MR imaging to be performed in all patients with a newly diagnosed breast cancer?

No, they are not. On this basis alone (or based on other similar studies), we cannot say yes. Fischer's article reports an observational retrospective study, not a prospective randomized clinical trial. The authors correctly report the characteristics of the two groups of patients: the MR group had 88% of invasive tumors and a 12% of in situ tumors while in the non-MR group these data were 96% and 4%, respectively. The tumors in stage pT1 were 64% in the MR group and only 48% in the non-MR group, those in stage pT3-4 7% and 28%, respectively. Moreover, the MR group had a higher number of patients with negative nodal status and lower histologic grading. Here the problem is that the two retrospective groups are not similar. The patients belonging to the MR group had smaller and less invasive and aggressive tumors. There is bias in favor of the MR group, at least for the rate of breast conserving surgery and ipsilateral recurrence. The result on the rate of contralateral cancers is probably more robust but does not have the sharpness which could be obtained by a randomized controlled trial. At any rate, for the supporters of preoperative breast MR imaging it is good news that a retrospective study did not reveal an increased rate of mastectomies, but the findings are not conclusive. A recent meta-analysis [Houssami et al, 2008] confirmed the therapeutic impact of preoperative breast MR imaging. But the evaluation of the outcome impact will be possible only when results from randomized studies will be available.

9.2.2. Subject Selection

Selection bias

Patient selection is a fundamental issue for any study. If you want to translate the results of a study to clinical practice, the study sample needs to be repre-

sentative of the population on which you want to apply these results. Bear in mind that some *selection bias* cannot be avoided. Sometimes it is produced by the context or related to the study design. At any rate, the authors must acknowledge this explicitly in the Discussion section, in order to avoid that the readers draw the wrong conclusions.

Centripetal bias

We name *centripetal bias* the distortion due to a high concentration of rare, difficult, or complex cases in a large, specialized, or university hospital.

Popularity bias

Popularity bias is the same effect obtained by the physicians who voluntarily selected these cases.

Diagnostic access bias

Conditions of limited access (*diagnostic access bias*) can be due to the geographic location of the hospital or to the social and economic level of the patients who could access it.

Referral bias

Centripetal bias, popularity bias, and diagnostic access bias can be grouped into what is known as *referral bias*.

Patient filtering bias

The enrollment can favor symptomatic or high-risk subjects or subjects with peculiar demographic features (*patient filtering bias*).

Diagnostic safety bias

Invasive diagnostic modalities or procedures associated with non-negligible risks (ionizing radiation exposure, contrast material injection) may have a higher likelihood of being performed in subjects suspected of having a disease (*diagnostic safety bias*).

Co-treatments

The presence of *co-treatments* (therapies, or other diagnostic examinations) can limit enrollment or determine changes in the radiologic findings in all the subjects of the sample (if co-treatments are administered only to a part of the sample, the internal validity of the study is flawed).

Spectrum bias

A well known bias in subject selection is the disease *spectrum bias* [Ransohoff and Feinstein, 1978]. This happens when disease type (e.g. histologic type), severity (e.g. tumor stage), or duration (e.g. acute or chronic phase) of the enrolled patients are clearly different from those of patients commonly treated in clinical practice.

Population bias

For spectrum bias we strictly mean an inhomogeneity in the characteristics of the disease in enrolled patients, with a relevant difference with those we find in clinical practice. A relatively different case is that of an unusual (too high or too low) disease prevalence in the studied sample in comparison with what we find in clinical practice: this case is named *population bias*.

Patient cohort bias

Of course, these two biases can sum each other, with a large negative effect on the external validity of the study. Spectrum bias and population bias can be grouped under the name of *patient cohort bias*.

A small sample of patients with advanced disease and healthy volunteers (huge spectrum bias)

There are particular situations where the study design knowingly implies a spectrum bias. This is the case of the first attempts to evaluate the diagnostic performance for a given disease of a new imaging modality. Suppose we plan an observational prospective transversal study: the new modality is performed on a small sample of patients with advanced, fully evident disease and on a small sample of healthy volunteers[1]. Now suppose that the new imaging modality is demonstrated unable to distinguish the patients with advanced disease from healthy volunteers. Without wasting time and money, we know that the

[1] Note that due to subject selection this study has the logical structure of a comparison between cases and controls. However, the chronologic features of a study like this, beginning with a known diagnosis (disease or no disease) and subsequently investigating the result of a future event (a radiologic examination) prevent us from classifying this as a classic case-control study (which instead investigates past events; see Section 8.2).

new modality is not useful at all for this application. Conversely, if the new imaging modality is demonstrated to make this distinction, we can go ahead with other larger studies. However, even in the event of a very good performance of the new modality in this first study, we cannot transfer this experience to clinical practice. The reason is the presence of *a huge spectrum bias*. With reference to Chapter 1 (Section 1.6, Figures 1.7 and 1.8), the two curves of the distribution of patients and healthy volunteers are markedly shifted away from each other along the x-axis with a consequent reduction in false positives and false negatives. In clinical practice, we instead find patients with less advanced disease, patients with a different disease but similar symptoms, non-volunteer healthy subjects who are older and with similar symptoms, and comorbidities in patients and healthy subjects.

Enthusiasm for the first reports, afterwards delusion

This is one of the reasons explaining the enthusiastic results of the first reports on the diagnostic performance of a new imaging modality or technique, which later are frequently downsized. An example of this initial overestimation of the diagnostic performance was described for the diagnosis of carpal tunnel syndrome with MR imaging [Radack et al, 1997].

At the other end of the scale, an alternative case of spectrum bias which is less common in the radiologic literature is to put a diagnostic modality on the benchmark of a series of cases selected for high diagnostic complexity. The obvious result is an underestimation of sensitivity and specificity.

Comorbidities

Lastly, *comorbidities* can differently characterize a study sample, influencing the diagnostic performance of a radiologic examination.

After the initial experimental phase, if our purpose is to estimate the diagnostic performance of an imaging modality to be applied in clinical practice, the solution to the problem of patient cohort bias is to enroll a random sample of the population which could undergo the imaging modality in clinical practice. If we are testing a screening method, we invite a random sample of asymptomatic subjects with suitable demographics (age and sex). If we are testing a method for clinical diagnosis, we enroll a random sample of patients with clinical features (symptoms, findings at physical examination or at previous exams) to have a defined pretest disease probability, certainly higher than that of the sample used to test a screening method.

Enrolling a consecutive series of patients

When we want to test a method for clinical diagnosis, the practical solution most frequently adopted is to try to enroll a *consecutive series of patients*. This solution however does not solve all the problems. In fact, even a consecutive series is the result of a selection done by the local context, due to a number of factors, such as:

– geographic location of the testing site;
– temporal interval of the test during the year (important for seasonal diseases);
– type of hospital or diagnostic center;
– selection of out-patients or in-patients (or ratio between them in the sample);
– selection by referring physicians.

Thus, all studies estimating the clinical (non-screening) diagnostic performance of a method are limited to some degree by patient cohort bias.

For all these reasons, the section of *Materials and Methods* of a paper reporting the results of a clinical study should provide:

Give complete information about the enrolled patients

1. detailed information regarding demographic and clinical data of the studied patients (with inclusion and exclusion criteria explicitly stated);
2. an accurate description of all the patients excluded up until the definition of the sample from which the results of the study are derived, with particular reference to:

   a. the refusals to be enrolled (listing the reasons);
   b. the cases of enrolled patients who did not perform the diagnostic examination (listing the reasons);
   c. the cases of non assessability of the examination (e.g. low image quality due to artifacts);
   d. the cases with a result non-classifiable as positive or negative (indeterminate results);
   e. the cases of non performed reference standard (listing the reasons);
   f. the cases of non assessable or non classifiable reference standard (listing the reasons).

9.2.3. Radiologic Methods and Reference Standard

The absence of bias due to the choice of radiologic methods and reference standard is a basic condition for achieving good study quality. It can be ascertained by analyzing the subsections of materials and methods named *Imaging Methods* and *Image Analysis* (where we find the radiologic expertise of the authors), commonly followed by *Pathologic Examination* or *Standard of Reference*. In some articles a specific subsection is dedicated to *Radiologic-pathologic Correlation*. From the viewpoint of possible bias, we can list *four radiologic factors* and the *reference standard*:

1. diagnostic technology;
2. radiologic technique (protocol for performing the examination);
3. methods for image interpretation;
4. training and specific experience of the readers;
5. reference standard.

9.2.3.1. Diagnostic Technology (Technologic Obsolescence)

Technologic obsolescence

The use of clearly outdated technology brings an obvious bias determining an underestimation of the diagnostic performance. For example, this criticism would affect a study on the diagnostic performance of CT for pulmonary embolism using single-slice (non multidetector) equipment, of ultrasound for breast cancer with a low-frequency transducer, of intracranial MR angiography with a low field magnet, and so on.

However, the high speed of technologic evolution makes this a nontrivial problem. We cannot limit ourselves to recommending the use of updated technology. Studies with clinical outcome end-points with prospective longitudinal design frequently plan *many years of follow-up*. When the results finally arrive, the technology used might already be outdated, reducing (sometimes negating)

the possibility of applying the results to clinical practice. The development of technology may be so rapid that clinical studies on diagnostic performance necessitating a large sample of patients (with relatively prolonged enrollment and also with a multicenter design) imply a high risk of being published when the technology used in the study is clearly out of date. An example of this risk is given by the subsequent generations of multidetector CT equipment from 1999 to the present time: from 4 to 8, 16, 32, 64, 256, 320 rows of detectors… And flat-panel CT is on its way…

When you design a study, use updated technology and take into consideration the time needed to complete patient enrollment in relation with the predictable (how predictable?) technologic evolution.

9.2.3.2. Imaging Protocol

Imaging protocol bias

Here the factors which could generate an error determining a bias are practically infinite: all kinds of technical imaging parameters; patient preparation and positioning; dosage and administration regimen of contrast material; timing and modality of image acquisition; postprocessing procedures, etc. An error in any of these (or other) factors can have a negative impact on the external validity of the study.

9.2.3.3. Imaging Analysis

Imaging analysis bias

Interpretation bias

The preliminary definition of the methods for imaging interpretation and analysis, including all measurements performed on them, is another crucial aspect of a study. If the interpreting method is incorrect, if the distinction between negative and positive findings is not entirely defined, if the methods of measurement of any imaging feature entering the results are not precisely defined, the external validity could be flawed.

9.2.3.4. Reader Training and Experience

Bias from reader training/experience

Readers with little experience using the specific imaging modality and technique might underestimate the diagnostic performance. Conversely, highly-experienced hyperspecialized radiologists (i.e. those working in a university center or in a hospital highly devoted to particular diseases) could overestimate the diagnostic performance.

This is another characteristic which renders multicenter studies, especially when conducted across multiple countries, clearly superior to single-center studies.

9.2.3.5. Reference Standard

Imperfect standard bias

Here we should consider the choice of the reference standard and not its application. We name *imperfect standard bias* the distortion due to the use of an inadequate reference standard. An example is the use of pulmonary angiogra-

phy as reference standard for the diagnosis of pulmonary embolism with multidetector row CT [Sica, 2006].

The general problem of the reference standard for studies on diagnostic performance is complex. We cannot always obtain a histopathologic evaluation of all the positive and negative findings of a diagnostic examination. In screening programs, the reference standard is given by a combination of histopathology of clearly positive and suspicious findings and clinical and imaging follow-up of the patients with negative examinations. Moreover, several imaging modalities supply in vivo functional information that cannot be verified with a reliable standard of reference. A typical example is the evaluation of cardiac end-diastolic and end-systolic volumes and of the ejection fraction. In similar cases, the imaging modality demonstrated to be the most reliable is adopted as a reference standard: for many years, echocardiography has been compared with cine MR imaging, the latter being characterized by a higher but surely not perfect intra- and interobserver reproducibility [Sardanelli et al, 2008].

The rule is to choose the best possible reference standard in relation to the sample under investigation and the population from which the sample is taken.

9.2.4. Statistical Analysis

Bias from statistical methods

The choice of the cutoff for α error and β error, which are the threshold for significance and the complement to 1 of the power of the study, respectively, are two very important features of study design. Apart from rare cases, α error is always put at the level of 0.05. On the other hand, as we saw in Chapter 8, an error in evaluating the difference between the samples considered to be clinically relevant can determine important errors in the sample size calculation and, as a consequence, in the study power. The absence of sample size calculation implies the risk of a false negative study.

Lastly, if you choose statistical tests which are not suitable to the study design, the type of measured variables or their distribution, false positive and false negative results are possible, with a complete failure of external validity.

9.3. Bias Affecting Internal Validity

Flaws in implementation limit the internal validity of a study on diagnostic performance because they undermine its intrinsic logical coherence. Four categories can be distinguished, as follows: protocol application, reference standard application, data measurements and reader independence.

9.3.1. Protocol Application

Protocol violations

In Section 9.1 we stated that a high number of *protocol violations* implies errors or feasibility problems and undervaluation in planning and designing the study. If the protocol is not respected, a study loses its internal validity.

9.3.2. Reference Standard Application

Verification bias

Work-up bias

In *applying the reference standard* three types of bias are possible. *Verification bias* is due to the application of the reference standard only to a part of the sample investigated, causing serious overestimation or underestimation of sensitivity and specificity. *Work-up bias* is a peculiar kind of verification bias which happens when the reference standard is only applied to cases with a positive disease diagnosis confirmed by the examination under investigation. This is the case when we cannot obtain a histopathologic reference standard for lesions clearly diagnosed as benign or for subjects who proved to be unaffected by any lesion. In the absence of at least one negative follow-up in these negative cases, we have:

1. an overestimated sensitivity caused by lack of information regarding potential false negatives;
2. a reduced possibility of evaluating specificity due to a lack of information regarding many true negatives.

Incorporation bias

Incorporation bias consists of the use of the result of the examination under investigation as one of the parameters determining the reference standard. An example is the use of the neurologic diagnosis at discharge as a reference standard for the study of diagnostic performance of CT and MR imaging in patients with acute stroke [Mullins et al, 2002]. In fact, this neurologic diagnosis is also based on the result of the two examinations for which it should work as reference standard.

9.3.3. Data Measurement

In data measurement, six types of bias are possible:

Disease progression bias

1. *disease progression* (too long a time interval between examinations and reference standard);

Drop-out bias

2. *withdrawals from the study* with lack of reference standard (drop out);

Indeterminate results

3. *indeterminate results* (typically due to technical artifacts). If the examinations can be repeated, they cause higher economic, and sometimes biologic, costs. If the examinations cannot be repeated, indeterminate results should be carefully counted and declared; according to the study design and protocol, they can be excluded or included (as false negatives or false positives in relation to the reference standard) from calculation of diagnostic performance;

Lost at follow-up data

4. *data lost at follow-up* (especially when we suspect that these lost subjects could have had a mean different result from that of patients who remained available at follow-up);

Intra- and inter-observer variability

5. *reader variability* (also the external validity of a study depends on intraobserver and interobserver reproducibility – see Chapter 7);

Temporal effects

6. *temporal effects* (due to a learning curve of the readers or to changing technology during a study).

Particular types of bias due to disease progression are those we face in screening programs without a control group: the so-called lead time bias and length

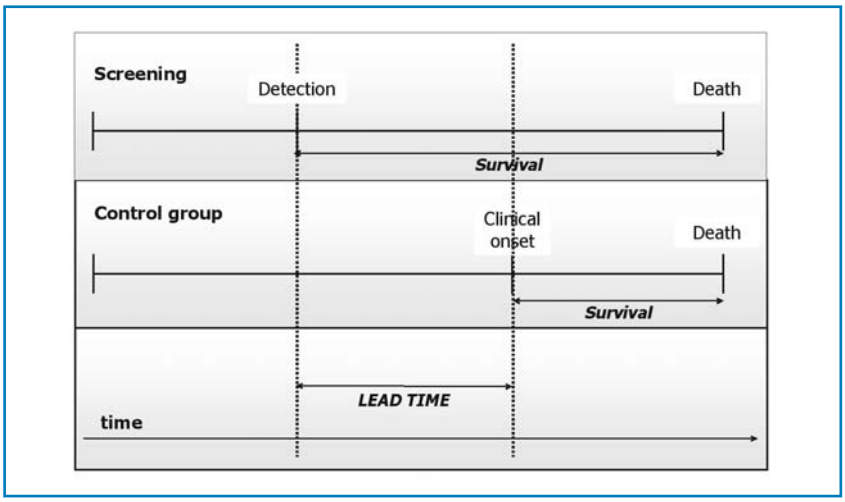


Figure 9.2. Lead time bias. Comparison between a group of screened subjects and a control group. The survival in the screened subjects appears double that of the controls. However, if the control group has been formed as a result of correct randomization (with the same probability of disease of the same mean severity as the screened group), the increase in survival is revealed as only apparent, solely due to earlier diagnosis. The difference between apparent and real survival in the screened group is the *lead time.*

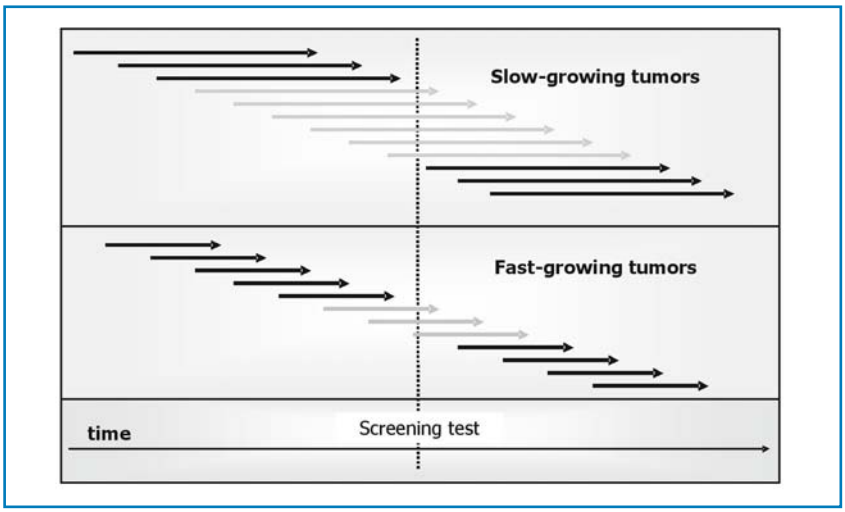


Figure 9.3. Length bias. A screening event, given a fixed time interval between one event and the next, can more probably reveal slow-growing tumors than fast-growing tumors. The extent of the arrows represents the time between the subclinical detectability and the clinical onset. Black arrows represent tumors not detected at the screening test (interval cancers) while gray arrows represent screen-detected tumors.

bias. In *lead time bias* the earlier diagnosis creates a false effect of prolonged survival of the group of subjects who were screened (Figure 9.2); in *length bias* a different disease progression acts making the *overdiagnosis* of indolent tumors more probable in the group of subjects who were screened (Figure 9.3)

Lead time bias

Length bias

9.3.4. Reader Independence

Four types of bias can limit the reading:

Diagnostic review bias
1. *diagnostic review bias:* the reference standard was defined by a person aware of the result of the examination under investigation;

Test review bias
2. *test review bias:* the result of the examination under investigation was defined by a reader aware of the result of the reference standard;

Comparator review bias
3. *comparator review bias:* the result of one of two examinations under investigation was defined by a reader aware of the result obtained with the other examination;

Clinical review bias
4. *clinical review bias:* the result of the examination under investigation was defined by a reader who knew demographic data and clinical status of each patient, a situation which is similar to clinical practice but potentially able to incorporate pretest probability of the disease in the result of the examination.

9.4. A Lot of Work to Be Done

We should be better

The demand for improving the quality of research on diagnostic performance was well shown in a study published in *JAMA* in 1995 [Reid et al, 1995]. The authors reviewed 112 articles regarding diagnostic tests published from 1978 to 1993 in four important medical journals. Overall, over 80% of the studies had relevant bias flawing their estimates of diagnostic performance. In particular:

– only 27% of the studies reported the disease spectrum of the patients;
– only 46% of the studies had no work-up bias;
– only 38% of the studies had no review bias;
– only 11% of the studies reported the confidence interval associated with the point estimates of sensitivity, specificity, predictive values etc.;
– only 22% of the studies reported the frequency of indeterminate results and how they were managed;
– only 23% of the studies reported the reproducibility of the results.

In this context, a detailed presentation of the rules to be respected for a good quality original article on diagnostic performance was outlined in a paper concerning the STARD initiative [Bossuyt et al, 2003]. It provides an extremely useful checklist of 25 items to be verified to avoid omitting important information. This checklist is entirely reproduced in the next chapter. The gap to be filled was testified by Smidt et al in a study published in 2005 [Smidt et al, 2005]. They evaluated 124 articles on diagnostic performance published in 12 journals with an impact factor of 4 or higher, using the STARD checklist. Only 41% of articles reported on more than 50% of STARD items, and no articles reported on more than 80%. A flow chart of the study was presented in only two articles. The mean number of reported STARD items was only 12. They concluded that "Quality of reporting in diagnostic accuracy articles published in 2000 is less than optimal, even in journals with high impact factor".

The relatively low quality of studies on diagnostic performance is a relevant threat to the successful implementation of evidence based radiology. Hopefully,

the adoption of the STARD requisites will improve the quality of radiologic studies but the process seems to be very slow [Hollingworth and Jarvik, 2007], as demonstrated also by the recent study by Wilczynski [Wilczynski, 2008].

A lot of work still remains to be done.

References

Brealey S, Scally AJ (2001) Bias in plain film reading performance studies. Br J Radiol 74:307-316

Bossuyt PM, Reitsma JB, Bruns DE et al (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Radiology 226:24-28

Fischer U, Zachariae O, Baum F et al (2004) The influence of preoperative MRI of the breasts on recurrence rate in patients with breast cancer. Eur Radiol 14:1725-1731

Hollingworth W, Jarvik JG (2007) Technology assessment in radiology: putting the evidence in evidence-based radiology. Radiology 244:31-38

Houssami N, Ciatto S, Macaskill P, et al (2008) Accuracy and surgical impact of magnetic resonance imaging in breast cancer staging: systematic review and meta-analysis in detection of multifocal and multicentric cancer. J CLin Oncol 26:3248-32

Kelly S, Berry E, Roderick P et al (1997) The identification of bias in studies of the diagnostic performance of imaging modalities. Br J Radiol 70:1028-1035

Lehman CD, Gatsonis C, Kuhl CK et al (2007) MRI evaluation of the contralateral breast in women with recently diagnosed breast cancer. N Engl J Med 356:1295-1303

Mullins ME, Schaefer PW, Sorensen AG, et al (2002) CR and conventional and diffusion-weighted MR imaging in acute stroke: study in 691 patients at presentation to the emergency department. Radiology 224:353-360

Peters NH, Borel Rinkes IH, Zuithoff NP et al (2008) Meta-analysis of MR imaging in the diagnosis of breast lesions. Radiology 246:116-124

Radack DM, Schweitzer ME, Taras J (1997) Carpal tunnel syndrome: are the MR findings a result of population selection bias? AJR Am J Roentgenol 169:1649-1653

Ransohoff DF, Feinstein AR (1978) Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 299:926-930

Reid MC, Lachs MS, Feinstein AR (1995) Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 274:645-651

Sardanelli F, Quarenghi M, Di Leo G et al (2008) Segmentation of cardiac cine MR images of left and right ventricles: interactive semiautomated methods and manual contouring by two readers with different education and experience. J Magn Reson Imaging 27:785-792

Sardanelli F, Podo F, D'Agnolo G et al (2007) Multicenter comparative multimodality surveillance of women at genetic-familial high risk for breast cancer (HIBCRIT study): interim results. Radiology 242:698-715

Sica GT (2006) Bias in research studies. Radiology 238:780-789

Smidt N, Rutjes AW, van der Windt DA et al (2005) Quality of reporting of diagnostic accuracy studies. Radiology 235:347-353

Wilczynski NL (2008) Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. Radiology 248:817-823

10

How to Write a Radiologic Paper

While we teach, we learn.

SENECA

Practical rules and recommendations

The purpose of this chapter is to provide a list of practical rules and recommendations for writing a scientific article, with particular reference to radiology. First, we will try to define the main types of articles published in the most important journals[1], with particular reference to *major papers* (composed of the four classic sections *Introduction*, *Materials and Methods*, *Results* and *Discussion*). Second, we will evaluate the radiologic journals with the recent trend of their *impact factor* (IF) – explaining its mechanism of calculation – compared with that of nonradiologic journals, a comparison useful for the choice of the suitable journal for submitting an article. Third, we will explain the absolute need of Ethics Committee approval and of informed consent by the patients asked to participate in a clinical study. Fourth, we will illustrate the content of each of the four sections of the major papers and the other associated sections, in particular the *Abstract* and the *References*. Moreover, we will provide several partial suggestions for tables, graphs, and figures, as well as some indications on how to interpret the Editor's response and the comments and criticisms of the reviewers.

A very important general reference for these topics is the article *Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication* by the *International Committee of Medical Journal Editors*, available as updated to October 2007 at http://www.icmje.org/index.html, where the reader can find further information.

[1] We basically refer to a series of radiologic journals with high impact factor (*Radiology*, *Invest Radiol*, *Eur Radiol*, *Magn Reson Med*, *J Magn Reson Imaging*, *AJNR Am J Neuroradiol*, *AJR Am J Roentgenol*) but the general content of this chapter is also valid for the other radiologic and nonradiologic journals.

F. Sardanelli, G. Di Leo, *Biostatistics for Radiologists*,

Some considerations and recommendations reported here can be found in the *instructions for authors* in the journals themselves. Several of them are the result of the personal experience of the senior author of this book, first as author and then also as reviewer for radiologic journals over a twenty year period, and are the product of a process of "trial and error" mechanism. You learn from your mistakes, and also from those of others.

10.1. Major Papers, Minor Papers, Invited Papers

Three main categories of papers

An outline of the types of articles published by radiologic journals (and also by nonradiologic journals) is presented in Table 10.1. We can distinguish three main categories: *major papers*, *minor papers* and *invited papers*.

Invited papers

Invited papers are articles requested by the editor from experts. They can be *editorials*, which frequently comment on major papers published in the same issue of the journal, or *narrative reviews*[2], namely critical summaries of published articles on an emerging clinical or research topic. Both are usually requested from authors who have already published important articles on the

Table 10.1. Types of papers

Major papers
Original articles (Original research) on humans
Experimental studies (on animals or phantoms)
Meta-analyses (systematic reviews)
Minor papers
Letters to the Editor
Brief communications (Preliminary reports)
Technical developments (Technical notes)
Case reports
Teaching articles
Pictorial reviews
Diagnosis please
Interpretation corner
Signs in imaging
Images in medicine
(...)
Invited papers
Editorials
Narrative Reviews
Position papers, Guidelines
Special reports
Special series

[2] Today review articles should be distinguished in *narrative reviews*, summaries of the literature on a given topic mostly reflecting the authors' opinion and experience, and *systematic reviews or meta-analysis* which should be considered as an original article including studies instead of patients (see Chapter 8). On the other hand, the meaning of the term review is entirely different for the so-called *pictorial reviews*, where authors present a spectrum of high quality images of a particular disease studied with one or more imaging techniques.

same topic. The instructions for authors can suggest the possibility of submitting a spontaneous review, but we advise against this if you have not already published some articles on the topic you would like to review.

A particular kind of invited papers are *position papers* and *guidelines*. They are basically official documents, edited by a panel of experts, commonly on behalf of the scientific committee of a medical society, which take a position on an emerging topic or define a guideline for the correct use of a diagnostic or therapeutic medical technology.

In recent years, new types of invited papers have appeared thanks to the presence of *special reports* and *special series*, commonly dedicated to the diffusion of particular knowledge by means of articles in subsequent issues of a journal. Examples of this in the journal "Radiology" are *Statistical Concept Series*, *Historical Perspectives*, and *What the Clinicians Want to Know*.

Major papers: original articles systematic reviews (meta-analyses)

However, the main target of a scientist remains the *major papers*. They can be *original articles* (also known as *original research*) on humans or on animal models or phantoms and *systematic reviews* or *meta-analyses* (see Chapter 8).

Having reached the end of this book, the reader will have understood that writing an original article for a medical journal is only the final act of a long process. In fact, before editing the text of an original article, the following phases should have been completed:

1. definition of the experimental hypothesis (H_1);
2. study design and definition of the null hypothesis (H_0);
3. sample size calculation;
4. editing of the study protocol;
5. request of approval from the Ethics Committee;
6. enrollment of patients (obtaining informed consent) and implementation of the study protocol;
7. data acquisition;
8. statistical and logical data analysis.

This is the flowchart for a typical study with a prospective design. It also basically works for a retrospective study, with the (partial) exception of the sample size calculation. If a study has been conducted according to this process, many problems in editing the paper have already been solved. *Introduction* and *materials and methods* sections have already been written, at least in a large part of them, for the documents submitted to the Ethics Committee. The logical scheme is not so different for a *systematic review* (see Chapter 8), with the advantage of not needing approval from the Ethics Committee. In the following paragraphs we basically refer to problems in writing and editing a major paper, in particular an *original article on humans*.

Minor papers

Lastly, there is the possibility of writing a *minor paper*.

The *letters to the editor* allow the presentation of short considerations on an article already published in the same journal or opinions on a particular topic. Sometimes the editorial space offered for a letter is very limited (up to only 400-500 words with no more than five references) and the interval time is restricted to a few weeks from the publication of the article you want to comment on. To avoid wasting your time, we advise you carefully read the paragraphs in the Instructions for Authors dedicated to this kind of article in the journal you have chosen.

Brief communications or *preliminary reports*, *technical developments* (where some journals place the *experimental studies on phantoms*) or *technical notes* are necessarily short articles with the same logical structure as the original articles: preliminary clinical observations on a small sample of patients; new technical procedures performed on phantoms, small samples of healthy volunteers or patients. Conceptually, they have the same difficulties as original articles. Sometimes, the absolute need to be brief creates more problems than those we meet with original articles, including a hard limitation in the number of references.

Case reports

Case reports deserve some special considerations. Over the past decades, these kinds of articles were a way to start scientific writing for several generations of radiologists. It was a good training phase for beginners. Today, things have changed. There are many reasons for advising against investing your time in writing case reports.

First, the following opinion – unquestionable from a methodologic viewpoint – has gained increasing favor: *the scientific value of anecdotal observations on single or few cases is low due to the impossibility of quantifying the probability of any described event*. As a consequence, a published case report is rarely quoted by other authors. A journal which publishes many case reports has a significant disadvantage in terms of impact factor. The reader can verify this tendency by means of the number of imaging journals which still accept the submission of case reports: only eighteen, mainly placed in the middle-low part of the ranking. A long waiting list before publication after acceptance is a frequent reason presented by the editor for rejecting an interesting case report. A higher probability of acceptance exists for a case report of a very rare disease.

Second, the journals still admitting the submission of case reports require that the article is well written, with detailed methods and updated references. Thus, the time needed for writing and editing such an article is not that much less than needed for an original article. It is therefore better to take the bull by the horns and choose to perform a real study and write an original article. The probability of acceptance is undoubtedly higher.

Teaching articles

Interesting clinical cases can be used for a different scientific purpose. In fact, several journals propose *teaching articles,* sometimes by means of supplements or additional publications (an example is *AJR Integrative Imaging*). These frequently offer the reader the possibility of gaining points for *continuous medical education*. A case report can be transformed into an educational article which regards an unusual diagnosis, describing a radiologic sign, showing high quality images (as it is for *pictorial reviews*), or summarizing the basics for a differential diagnosis.

10.2. Which Medical Journal?

Each journal has special rules

Before starting to write you should have decided to which journal you want to submit the manuscript. In fact, even though the logical structure is the same, *each journal has its own formal and stylistic rules*[3] *which need to be accurate-*

[3] By stylistic rules we mean here the guidelines regarding the type of Abstract and its subtitles, the titles of sections and subsections, the format of the references, not the style of printing.

ly respected. To avoid problems and wasting time, you should know them from the beginning.

Congress abstract and full papers

The first draft of the results of a study is commonly an abstract for a congress, a summary produced within the congress deadline for submission. We might suggest a more virtuous mechanism: submitting the congress abstract at the same time as the submission of the full paper to the journal (maybe the journal of the medical society organizing the congress). However, especially for those colleagues who must subdivide their time between clinical and scientific work, the congress abstract submission deadlines are useful catalysts which prompt producing the results of ongoing studies.

The acceptance of a congress abstract is obviously a positive event, indicating content judged to be original or at any rate interesting by the congress reviewers. However, neither is it an open road to acceptance of the related paper in the associated journal, nor a mandatory requirement for submitting the paper to the journal. In practice, you can submit a paper to the journal even though you did not submit the abstract to the congress or you submitted the abstract but it was not accepted. As a result of practical experience, considering the Annual Meeting of the Radiological Society of North America (RSNA) and the journal *Radiology* or the European Congress of Radiology and the journal *European Radiology*, all the four combinations of events are possible:

1. a work accepted as abstract at the congress was accepted as full paper for publication in the journal;
2. a work rejected as abstract at the congress was rejected by the journal;
3. a work accepted as abstract at the congress was rejected by the journal;
4. a work rejected as abstract at the congress was accepted as full paper for publication in the journal.

Combinations 3 and 4 should not astonish the reader, for at least three reasons. First, the reviewers of a congress abstract and those of the paper are almost always different persons and their opinions can differ highly. Second, the full paper makes positive and negative aspects of the study clearly visible. Third, with reference to combination 4, when writing the full paper you may conduct broader data analysis, and when doing so discover new aspects especially for data interpretation and comment.

To which journal do you submit the manuscript?

Apart from the relation between congresses and journals, which journal should we submit a paper to? Generally, we advise trying a highly ranked journal first. Bear in mind that the review process is normally characterized by double blindness: you do not know who the reviewers are and the reviewers do not know who you are. The reviewers' comments and criticisms often reveal real limitations to your study and frequently indicate interesting solutions. Therefore, even if the paper is rejected, you could obtain useful suggestions for submission to another journal. It is not a rare event that a paper rejected by a journal with a middle ranked impact factor is accepted by a journal with a higher impact factor, maybe after some modifications suggested by the reviewers of the first journal. *At any rate, the interaction with the reviewers of highly ranked journals is a real scientific school. You will learn from your errors much more than you think.* This is also true for statistical methods (e.g. one of the reviewers of *Radiology* is a statistician).

Impact factor (IF)

In this reasoning we have considered the IF – annually presented by the *ISI-Thomson Scientific* in the *Journal Citation Reports* (JCR) – as a reliable measure of the scientific value of a medical journal. The underlying hypothesis is that the number of quotations of a journal is directly proportional to its diffusion in the medical scientific community and that this diffusion is an indicator of its scientific level. For example, the 2007 IF of a journal is calculated according to the following formula:

$$2007\ \text{IF} = \frac{\text{Number of citations during 2007 of articles published by the journal in 2005 and 2006}}{\text{Number of articles published by the journal in 2005 and 2006}}$$

Technical limitations of the impact factor

We cannot develop here a deep analysis of the IF. Like many tools, it has *technical limitations*: exclusion of journals not written in English; inclusion in calculation of self-citations (when the journal quotes itself); competitive advantage in publishing more small articles than less large articles. However,

The basic limitation of the impact factor

the most important limitation is in the basic hypothesis that the scientific quality of a journal (or also an article or an author) can be obtained by means of a purely quantitative calculation.

At any rate, notwithstanding these limitations, the IF remains the only really available tool for concisely evaluating *the weight* of a journal. Obviously, given the large IF range obtained by the journals of different scientific medical fields, methods for *normalization* based on the rank of the IF of each journal within its medical field are frequently used to make comparisons between different fields possible. At the University of Milan School of Medicine, we use a method based on the subdivision into quartiles of the distribution of the IFs of journals within their medical field, giving a standardized score according to the quartile: 6 for the highest quartile, 4 for the second quartile, 2 for the third quartile, and 1 for the lowest quartile.

In Table 10.2 we report the original (non-normalized) IFs obtained in 2000, 2003, and 2006 (related to 1998-99, 2001-2002, and 2004-2005, respectively) of the journals regarding "radiology, nuclear medicine, and medical imaging".

A relevant aspect of the recent trend of IFs is a progressive increase in this mean index for the imaging and radiologic journals. No journal has an IF higher than 6.0, but many of them show an increasing trend from 2000 to 2006. In these years, the mean IF of imaging and radiologic journals went up from 1.469 to 2.053. This reflects the increasing role of imaging and radiologic techniques in clinical medicine and the improved quality of scientific production by radiologists and other colleagues working on medical imaging. In

European Radiology

this context, *European Radiology* went up from 1.321 in 2000 to 2.554 in 2006 and to 3.405 in 2007. By the way, in 2007 the official journal of the

La Radiologia Medica

Italian Society of Medical Radiology, *La Radiologia Medica* (*Radiol Med*) obtained its first IF, ranked at 0.967, an important and promising result for Italian radiology. In Table 10.2 you can observe how the number of imaging journals with an increasing 2006 IF in comparison with both 2000 and 2003 is 57 out of 89 (64%).

The rules for editing a scientific article are in part the same as those for writing good scientific international English. Grammar and style rules (e.g.

Table 10.2. Impact factor (IF) of the journals regarding "radiology, nuclear medicine and medical imaging" obtained in 2000, 2003, and 2006

Journal	IF 2000	IF 2003	IF 2006		
Semin Radiat Oncol	2.427	3.604	5.889	*	
Neuroimage	6.857	6.192	5.559		
Radiology	4.130	4.815	5.251	*	#
J Nucl Med	3.617	4.899	4.986	*	
Hum Brain Mapping	5.163	6.058	4.888		#
Semin Nucl Med	2.143	3.431	4.473	*	
Int J Radiat Oncol	3.058	4.285	4.463	*	
Eur J Nucl Med Mol	-	3.324	4.041	*	
Radiother Oncol	2.469	2.870	3.970	*	
IEEE T Med Imaging	2.573	3.755	3.757	*	
Mean of the top ten	*3.604*	*4.323*	*4.728*		
Strhalenther Onkol	2.846	2.634	3.682	*	
NMR Biomed	1.914	3.333	3.626	*	
Med Phys	2.428	2.305	3.571	*	
Magn Reson Med	3.121	3.313	3.427	*	
Invest Radiol	1.410	1.990	3.398	*	
Med Imaging Anal	-	-	3.256	*	
Mol Imaging Biol	-	-	2.961	*	
Phys Med Biol	2.013	2.128	2.873	*	
J Biomed Opt	-	3.541	2.870	*	
J Magn Reson Imaging	-	2.694	2.637		
Radiat Res	2.752	3.208	2.602		
Eur Radiol	1.119	1.969	2.554	*	
Radiol Clin N Am	1.529	1.759	2.533	*	
J Nucl Cardiol	1.854	1.629	2.440	*	
J Vasc Interv Radiol	1.729	2.212	2.398	*	
Radiographics	1.396	2.063	2.344	*	
AJNR Am J Neuroradiol	2.126	2.629	2.279		#
Ultrasound Obst Gyn	1.725	1.973	2.288	*	
Clin Nucl Med	0.399	0.737	2.217	*	
Nucl Med Biol	1.580	2.000	2.121	*	
AJR Am J Roentgenol	1.863	2.474	2.117		
Ultraschall Med	0.925	1.473	2.103	*	
Q J Nucl Med	1.910	2.222	2.062		
Ultrasound Med Biol	1.822	2.033	2.011		
Nuklearmed-Nucl Med	0.965	1.849	1.990	*	
ROFO	1.005	1.786	1.976	*	
Concept Magn Reson A	-	-	1.872	*	
Int J Hyperther	0.952	1.762	1.866	*	
Acad Radiol	0.912	1.409	1.781	*	
Cancer Biother Radio	0.989	1.841	1.763		
J Cardiov Magn Reson	2.304	1.125	1.739		
J Radiat Res	1.111	1.697	1.709	*	
Clin Radiol	0.934	1.270	1.665	*	#
Neuroradiology	0.997	1.213	1.625	*	
Ultrasonic Imaging	1.794	1.576	1.606		
Magn Reson Imaging	1.452	1.420	1.580	*	#
J Comput Assist Tomogr	1.484	1.318	1.530	*	#
Magn Reson Mater Phy	-	1.836	1.514		
Korean J Radiol	-	1.783	1.483	*	
Brain Topogr	1.596	1.820	1.415		
Abdom Imaging	0.866	0.996	1.336	*	#
Eur J Radiol	0.822	1.060	1.332	*	#

(continued)

(continued)

Journal	IF 2000	IF 2003	IF 2006		
J Thorac Imag	0.663	0.923	1.328	*	#
Ultrasonics	0.711	0.780	1.322	*	
Int J Radiat Biol	2.586	2.165	1.312		
J Digit Imaging	0.722	0.953	1.304	*	#
J Neuroimaging	0.942	0.927	1.298	*	#
Nucl Med Commun	1.039	1.230	1.283	*	
Br J Radiol	0.951	1.089	1.279	*	#
J Ultras Med	0.966	1.194	1.189		#
Skeletal Radiol	0.695	0.821	1.176	*	#
Concept Magnetic Res	-	1.161	-		
Cardiovasc Intervent Rad	1.029	1.207	1.149		
Semin Ultrasound CT	0.797	0.851	1.135	*	
Int J Cardiovasc Imaging	-	0.496	1.119	*	#
Radiat Environ Bioph	1.110	1.131	1.090		
Pediatr Radiol	0.684	0.942	1.076	*	#
Appl Radiat Isotopes	0.716	0.690	0.924	*	
Comput Med Imaging Graph	0.500	1.158	0.909		
Neuroimag Clin N Am	1.095	0.663	0.905		
Health Phys	0.988	0.777	0.902		
Acta Radiol	0.785	1.096	0.884		
Dentomaxillofac Rad	0.780	0.669	0.821	*	
Ann Nucl Med	-	0.745	0.779	*	#
Clin Imaging	0.368	0.658	0.758	*	#
J Radiol Prot	-	-	0.736	*	
Radiologe	0.608	0.626	0.696	*	
Semin Roentgenol	0.597	0.887	0.625		
J Radiol	0.345	-	0.600	*	
J Clin Ultrasound	0.994	0.746	0.573		#
J Neuroradiol	0.451	0.603	0.509		
Radiat Prot Dosim	0.581	0.617	0.446		
Surg Radiol Anat	0.314	0.307	0.443	*	
Interv Neuroradiol	0.585	0.512	0.366		
Can Assoc Radiol J	0.268	0.376	-		
Riv Neuroradiol	0.051	0.152	-		
Int J Neuroradiol	0.139	-	-		
Mean	**1.469**	**1.808**	**2.053**		

Decreasing order according to the 2006 IF. The asterisk (*) indicates the journals with a 2006 IF increasing in comparison with both that of 2000 and that of 2003. The hash (#) indicates the journals which have specific indications for case report submission in their instructions for authors, accessed online from February 20 to March 8, 2008.
Source of IFs: Journal Citation Reports™ – Science edition, published by Thomson Scientific (with permission).

the use of tense or of active instead of passive verb forms, the use of upper-case or lower-case letters, etc.) go together with the technical rules for the units of measurement or for rounding numbers in the data, etc. A useful tool for this problem is the manual by Sylvia Rogers [Rogers, 2007]. Beginners can also refer to recently published articles in the journal to which they want to submit their paper.

If writing in English poses no problems (maybe with the cooperation of a more experienced colleague), you may dare to submit your paper to a journal ranked in the first IF quartile. Do not lose heart if the journal rejects your paper.

A paper which appears in a journal of the first IF quartile might have been previously rejected by two other journals.

To publish on non-radiologic journals

Publishing in nonradiologic journals is a relatively different task. Of course, only radiologic studies of general interest can be published in the big medical journals, such as *New Engl J Med, Lancet*, *JAMA*, *Ann Intern Med*, or other journals ranked with a very high IF. The access to journals with a limited medical field is easier. Here you can be pleasantly surprised: a paper rejected by a radiologic journal can be accepted by a clinical journal with a higher IF. As the scientific level of radiologic journals is increasing and the filtering by editors and referees is increasingly more stringent, you might consider the alternative of submitting it to a clinical journal if your paper has an evident clinical interest. The IF development of clinical journals with IF higher than 10 in 2006 is reported in Table 10.3. You can observe that the number of journals with an increasing IF is 70 out of 107 (65%), a percentage similar to that of imaging and radiologic journals. Obviously, the number of nonradiologic journals that can publish radiologic papers is much higher and comprises many medical journals with an IF lower than 10. The 2000, 2003, and 2006 IFs of the first ten journals of a series of medical fields is reported in Table 10.4.

Table 10.3. Journals with impact factor (IF) higher than 10 in 2006 and comparison with 2003 and 2000 (the asterisk indicates the journals with an increasing 2006 IF compared with 2000 and 2003)

	Journal	IF 2000	IF 2003	IF 2006
*	CA Cancer J Clin	24.674	33.056	63.342
*	New England J Med	29.512	34.833	51.296
	Annu Rev Immunol	50.340	52.280	47.237
	Annu Rev Biochem	43.429	37.647	36.525
*	Rev Mod Phys	12.774	28.172	33.508
*	Nat Rev Cancer	-	33.954	31.583
*	Physiol Rev	27.677	36.831	31.441
	Nat Rev Mol Cell Biol	-	35.041	31.354
*	Science	23.872	29.781	30.028
	Cell	32.440	26.626	29.194
*	Nat Rev Immunol	-	26.957	28.697
	Nat Med	27.905	30.550	28.588
	Annu Rev Neurosci	26.676	30.167	28.533
	Nat Immunol	-	28.180	27.596
	Nature	25.814	30.979	26.681
*	Annu Rev Cell Dev Bi	26.300	22.638	26.576
*	Chem Rev	20.036	21.036	26.054
*	Lancet	10.232	18.316	25.800
*	Brief Bioinform	-	-	24.370
	Nat Genet	30.910	26.494	24.176
*	Cancer Cell	-	18.913	24.077
*	Endocr Rev	19.524	17.324	23.901
*	JAMA	15.402	21.455	23.175
	Nat Rev Neurosci	-	27.007	23.054
	Nat Rev Genet	-	25.664	22.947
*	Annu Rev Pharmacol	19.289	21.786	22.808
*	Nat Biotechnol	11.542	17.721	22.672
*	Nat Rev Drug Discov	-	17.732	20.970

(continued)

(continued)

	Journal	IF 2000	IF 2003	IF 2006
*	Annu Rev Plant Biol	-	15.615	19.837
*	Nat Mater	-	10.778	19.194
*	Annu Rev Genet	13.450	11.920	19.098
	Nat Cell Biol	11.939	20.268	18.485
	Immunity	21.083	16.016	18.306
*	Mat Sci Eng R	6.083	-	17.731
*	Accounts Chem Res	13.262	15.000	17.113
*	Annu Rev Bioph Biom	16.194	13.351	16.921
*	Annu Rev Astron Astr	14.000	16.000	16.914
	Pharmacol Rev	25.381	27.067	16.854
*	Cell Metab	-	-	16.710
	Microbiol Mol Biol R	20.639	14.340	15.864
*	Nat Rev Microbiol	-	-	15.845
*	J Clin Invest	12.015	14.307	15.754
	Annu Rev Physiol	18.848	18.591	15.356
*	J Natl Cancer Inst	14.159	13.844	15.271
	Gene Dev	19.676	17.013	15.050
*	Behav Brain Sci	14.250	10.625	14.964
*	Nat Methods	-	-	14.959
*	Prog Polym Sci	3.698	7.759	14.818
	Nat Neurosci	12.636	15.141	14.805
*	Ann Intern Med	9.833	12.427	14.780
*	Annu Rev Microbiol	9.238	12.105	14.553
	J Exp Med	15.236	15.302	14.484
	Curr Opin Cell Biol	22.754	18.176	14.299
	Trends Ecol Evol	22.754	12.449	14.125
*	Plos Biol	-	-	14.101
	Mol Cell	18.195	16.835	14.033
*	Arch Gen Psychiat	11.778	10.519	13.936
	Neuron	15.081	14.109	13.894
	Trends Biochem Sci	13.246	14.273	13.863
*	Plos Med	-	-	13.750
*	Chem Soc Rev	10.747	9.569	13.690
*	Astron Astrophys Rev	3.455	3.600	13.667
*	J Clin Oncol	8.773	10.864	13.598
	Dev Cell	-	14.807	13.523
	Trends Neurosci	17.417	12.631	13.494
*	Annu Rev Med	9.891	11.381	13.237
*	Psychol Bull	6.913	8.405	12.725
*	Clin Microbiol Rev	12.141	11.530	12.643
*	Am J Hum Genet	10.351	11.602	12.629
*	Annu Rev Fluid Mech	6.486	5.108	12.469
	Gastroenterology	12.246	12.718	12.457
	Trends Cell Biol	18.815	19.612	12.429
*	Nat Chem Biol	-	-	12.409
*	Prog Lipid Res	5.379	10.000	12.235
*	Nat Phys	-	-	12.040
*	Lancet Infect Dis	-	-	11.808
*	Mol Psychiatr	8.927	5.539	11.804
*	Annu Rev Psychol	5.851	9.896	11.706
*	Cytokine Growth F R	6.049	9.600	11.549
*	Front Neuroendocrinol	8.375	8.870	11.526
*	Nat Struct Mol Biol	-	-	11.502
	Prog Neurobiol	9.933	12.327	11.304
*	Adv Catal	11.000	7.889	11.250
*	Annu Rev Phys Chem	9.237	10.500	11.250

(continued)

(continued)

	Journal	IF 2000	IF 2003	IF 2006
*	Curr Opin Struc Biol	10.427	8.686	11.215
	Curr Biol	8.393	11.910	10.988
*	Mass Spectrom Rev	7.600	7.364	10.947
	Circulation	10.893	11.164	10.940
	Annu Rev Genom Hum G	-	12.200	10.771
*	Immunol Rev	5.961	7.052	10.758
*	Aldrichim Acta	5.900	7.077	10.692
*	Adv Cancer Res	21.680	7.938	10.682
*	Annu Rev Biomed Eng	-	7.875	10.533
*	Annu Rev Nutr	7.071	9.326	10.449
*	Hepatology	7.304	9.503	10.446
	Phys Rep	7.110	11.980	10.438
*	Annu Rev Mater Res	-	5.333	10.400
	Trends Pharmacol Sci	10.377	13.965	10.400
*	Blood	8.977	10.120	10.370
*	Genome Res	7.615	9.635	10.256
*	Angew Chem Int Edit	8.547	8.427	10.232
	Progr Mater Sci	4.667	12.000	10.229
	Trends Immunol	-	18.153	10.213
*	Curr Opin Plat Biol	7.347	8.945	10.182
	J Cell Biol	13.955	12.023	10.152
*	Lancet Oncol	-	7.411	10.119
	Embo J	13.999	10.456	10.086
	Curr Opin Genet Dev	13.810	13.143	10.006
*	Semin Immunol	6.544	5.964	10.000
	Mean	**14.746**	**16.215**	**17.434**

Decreasing order according to the 2006 IF.
Source: Journal Citation Reports™ – Science edition, published by Thomson Scientific (with permission).

Table 10.4. List of the first ten journals of a series of medical fields (increasing order according to the 2006 IF and comparison with 2000 and 2003)

Sector / Journal	IF 2000	IF 2003	IF 2006
Allergy			
J Allergy Clin Immun	4.179	6.831	8.829
Allergy	2.385	3.161	5.334
Clin Exp Allergy	2.947	3.176	3.668
Immunol Allergy Clin	0.520	0.731	3.178
Pediatr Allergy Immu	1.635	1.573	2.849
Int Arch Allergy Imm	1.630	2.000	2.524
Contact Dermatitis	0.675	1.095	2.446
Ann Alerg Asthma Im	1.889	2.181	2.254
Curr Allergy Asthm R	-	-	2.016
Clin Rev Allerg Immu	0.741	1.173	1.677
Mean	**1.845**	**2.436**	**3.478**
Anatomy and Morphology			
Dev Dynam	3.131	3.160	3.169
J Anat	1.385	2.072	2.458
Anat Rec Part A	-	-	1.973
Cells Tissues Organs	0.896	1.757	1.841
Microsc Res Techniq	1.746	2.307	1.680

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
Appl Immunohisto M M	0.747	1.500	1.621
J Morphol	0.911	1.629	1.553
Adv Anat Embryol Cel	2.933	0.321	1.429
Anat Embryol	1.851	1.559	1.277
Zoomorphology	1.000	1.156	1.211
Mean	***1.622***	***1.718***	***1.821***
Andrology			
Int J Androl	1.357	1.588	2.183
J Androl	2.106	2.480	2.137
Asian J Androl	-	1.064	1.737
Andrologia	0.871	0.939	1.025
Arch Andrology	0.727	0.667	0.687
Mean	***1.265***	***1.348***	***1.554***
Anesthesiology			
Pain 3.853	4.556	4.836	
Anesthesiology	3.439	3.503	4.207
Euro J Pain	-	1.770	3.333
Brit J Anaesth	1.989	2.365	2.679
Clin J Pain	1.900	2.080	2.448
Anaesthesia	2.027	2.041	2.427
Anesth Analg	2.321	2.210	2.131
Region Anesth Pain M	1.129	1.766	2.056
Can J Anaesth	1.149	1.200	1.976
J Neurosurg Anesth	0.937	0.959	1.926
Mean	***2.083***	***2.245***	***2.802***
Cardiac and Cardiovascular Systems			
Circulation	10.893	11.164	10.940
Circ Res	9.193	10.117	9.854
J Am Coll Cardiol	7.082	7.599	9.701
Eur Heart J	3.840	5.997	7.286
Cardiovasc Res	3.783	5.164	5.826
J Mol Cell Cardiol	3.383	4.954	4.859
Trends Cardiovas Med	2.879	4.517	4.724
Basic Res Cardiol	1.490	2.993	3.798
Heart Rhythm	-	-	3.777
Am J Physiol-Heart C	3.243	3.658	3.724
Mean	***5.087***	***6.240***	***6.449***
Clinical Neurology			
Lancet Neurol	-	3.070	9.479
Ann Neurol	8.480	7.717	8.051
Brain	7.303	7.967	7.617
Cephalalgia	2.391	2.985	6.049
Neuroscientist	1.918	2.822	5.710
Neurology	4.781	5.678	5.690
Stroke	6.008	5.233	5.391
Brain Pathol	6.435	3.838	5.274
Curr Opin Neurol	3.176	3.920	5.229
Arch Neurol-Chicago	4.393	4.684	5.204
Mean	***4.987***	***4.791***	***6.369***
Critical Care Medicine			
Am J Resp Crit Care	5.443	8.876	9.091
Crit Care Med	3.824	4.195	6.599

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
Intens Care Med	2.098	2.971	4.406
J Neurotraum	2.877	2.587	3.453
Shock	2.785	2.542	3.318
Crit Care	-	1.911	3.116
Resuscitation	1.760	1.375	2.314
J Trauma	1.498	1.429	2.035
Crit Care Clin	-	1.485	1.845
Am J Crit Care	-	-	1.685
Mean	***2.898***	***3.041***	***3.786***
Emergency Medicine			
Ann Emerg Med	2.183	2.640	3.120
Resuscitation	1.760	1.375	2.314
J Burn Care Rehabil	0.810	1.042	1.744
Acad Emerg Med	1.419	1.844	1.741
Am J Emerg Med	1.054	1.489	1.518
Injury	0.363	0.511	1.067
Emerg Med J	-	0.633	0.869
J Emerg Med	-	0.652	0.816
Pediatr Emerg Care	0.428	0.505	0.700
Emerg Med Clin N Am	0.635	0.676	0.672
Mean	***1.082***	***1.137***	***1.456***
Endocrinology and Metabolism			
Endocr Rev	19.524	17.324	23.901
Cell Metab	-	-	16.710
Front Neuroendocrin	8.375	8.870	11.526
Recent Prog Horm Res	5.306	8.275	9.263
Diabetes	7.715	8.298	7.955
Diabetes Care	4.992	7.501	7.912
Trends Endocrin Met	3.908	7.850	7.066
J Bone Miner Res	5.877	6.225	6.635
J Clin Endocr Metab	5.447	5.873	5.799
Curr Opin Lipidol	5.661	6.966	5.689
Mean	***7.432***	***8.576***	***10.246***
Gastroenterology and Hepatology			
Gastroenterology	12.246	12.718	12.457
Hepatology	7.304	9.503	10.446
Gut	5.386	5.883	9.002
J Hepatol	3.761	5.283	6.073
Am J Gastroenterol	2.834	4.172	5.608
Semin Liver Dis	6.012	6.524	5.302
Gastrointest Endosc	2.820	3.328	4.825
Liver Transplant	2.130	4.242	4.629
Inflamm Bowel Dis	1.791	3.023	3.912
Am J Physiol-Gastr L	3.115	3.421	3.681
Mean	***4.740***	***5.810***	***6.594***
Genetics and Heredity			
Nat Genet	30.910	26.494	24.176
Nat Rev Genet	-	25.664	22.947
Annu Rev Genet	13.450	11.920	19.098
Gene Dev	19.676	17.013	15.050
Trends Ecol Evol	8.765	12.449	14.125
Am J Hum Genet	10.351	11.602	12.629

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
Annu Rev Genom Hum G	-	12.200	10.771
Genome Res	7.615	9.635	10.256
Curr Opin Genet Dev	13.810	13.143	10.006
Trends Genet	12.912	12.016	9.950
Mean	***14.686***	***15.214***	***14.901***
Geriatrics and Gerontology			
Rejuv Res	-	-	8.353
Aging Cell	-	-	6.276
Neurobiol Aging	4.159	5.552	5.599
Ageing Res Rev	-	3.795	4.526
Mech Ageing Dev	1.897	3.214	3.846
J Am Geriatr Soc	3.136	2.835	3.331
Age	2.622	-	3.034
Exp Gerontol	2.622	2.857	2.930
Am J Geriat Psychiat	-	3.741	2.894
J Gerontol A-Biol	1.549	4.369	2.861
Mean	***2.644***	***3.766***	***4.365***
Health Care Sciences and Services			
Milbank Q	4.568	3.524	6.794
Health Technol Asses	-	-	5.290
Med Care	2.535	3.152	3.745
Health Affair	3.823	3.673	3.680
Value Health	-	-	3.433
J Med Internet Res	-	-	2.888
Acad Med	1.554	1.104	2.607
Med Educ	1.078	1.188	2.467
J Pain Symptom Manag	-	1.885	2.437
Qual Saf Health Care	-	1.760	2.382
Mean	***2.712***	***2.327***	***3.572***
Hematology			
Circulation	10.893	11.164	10.940
Blood	8.977	10.120	10.370
Circ Res	9.193	10.117	9.854
Stem Cells	2.989	5.802	7.924
Arterioscl Throm Vas	5.111	6.791	6.883
Leukemia	3.736	5.116	6.146
Blood Rev	2.689	2.241	5.756
Curr Opin Hematol	-	4.449	5.202
J Thromb Haemost	-	-	5.138
Haematol-Hematol J	-	-	5.032
Mean	***6.227***	***6.975***	***7.325***
Immunology			
Annu Rev Immunol	50.340	52.280	47.237
Nat Rev Immunol	-	26.957	28.697
Nat Immunol	-	28.180	27.596
Immunity	21.083	16.016	18.306
J Exp Med	15.236	15.302	14.484
Immunol Rev	5.961	7.052	10.758
Trends Immunol	-	18.153	10.213
Semin Immunol	6.544	5.964	10.000
Curr Opin Immunol	12.549	12.118	9.422
J Allergy Clin Immun	4.179	6.831	8.829
Mean	***16.556***	***18.885***	***18.554***

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
Infectious Diseases			
Lancet Infect Dis	-	-	11.808
Clin Infect Dis	2.972	5.393	6.186
AIDS	8.018	5.521	5.632
J Infect Dis	4.988	4.481	5.363
Emerg Infect Dis	4.907	5.340	5.094
Antivir Ther	4.510	5.932	4.982
Curr Opin Infect Dis	0.778	2.674	4.795
AIDS Rev	-	-	4.022
Infect Immun	4.204	3.875	4.004
JAIDS-J Acq Imm Def	-	3.681	3.946
Mean	***4.340***	***4.612***	***5.583***
Medical Informatics			
J Am Med Inform Assn	3.089	2.510	3.979
J Med Internet Res	-	-	2.888
J Biomed Inform	-	0.855	2.346
Stat Med	1.717	1.134	1.737
Med Decis Making	2.152	1.718	1.736
Int J Med Inform	0.699	1.178	1.726
Method Inform Med	0.929	1.417	1.684
Artif Intell Med	1.793	1.222	1.634
IEEE T Inf Technol B	-	1.274	1.542
Stat Methods Med Res	-	1.857	1.377
Mean	***1.730***	***1.457***	***2.065***
Medical Laboratory Technology			
Crit Rev Cl Lab Sci	3.357	3.136	6.138
Clin Chem	4.261	5.538	5.454
Ther Drug Monit	-	2.372	3.032
Adv Clin Chem	1.600	0.917	2.440
Clin Biochem	1.327	1.825	2.331
Clin Chim Acta	-	1.633	2.328
Cytom Part B-Clin Cy	-	-	2.065
Clin Diagn Lab Immun	-	-	1.988
Clin Lab Med	0.460	0.854	1.904
J Lab Clin Med	1.978	2.011	1.812
Mean	***2.164***	***2.286***	***2.949***
Medicine, General and Internal			
New England J Med	29.512	34.833	51.296
Lancet	10.232	18.316	25.800
JAMA-J Am Med Assoc	15.402	21.455	23.175
Ann Intern Med	9.833	12.427	14.780
Plos Med	-	-	13.750
Annu Rev Med	9.891	11.381	13.237
Brit Med J	5.331	7.209	9.245
Arch Intern Med	6.055	6.758	7.920
Can Med Assoc J	2.352	4.783	6.862
Medicine	4.623	4.500	5.167
Mean	***10.359***	***13.518***	***17.123***
Medicine Research and Experimental			
Nat Med	27.905	30.550	28.588
J Clin Invest	12.015	14.307	15.754
J Exp Med	15.236	15.302	14.484

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
J Cell Mol Med	-	-	6.555
Trends Mol Med	-	-	5.864
Mol Ther	-	6.125	5.841
J Mol Med-JMM	3.445	4.101	5.157
Curr Mol Med	-	-	4.850
Gene Ther	5.964	5.293	4.782
Hum Gene Ther	6.796	4.965	4.514
Mean	***11.894***	***12.757***	***9.639***
Neuroimaging			
Neuroimage	6.857	6.192	5.559
Hum Brain Mapp	5.163	6.058	4.888
Psychiat Res-Neuroim	1.919	2.551	2.755
Cognitive Brain Res	2.733	2.865	2.568
Am J Neuroradiol	2.126	2.629	2.279
Neuroradiology	0.997	1.213	1.625
J Neuroimaging	0.942	0.927	1.298
Clin EEG Neurosci	-	-	1.255
Stereot Funct Neuros	-	0.425	1.195
Minim Invas Neurosur	0.805	0.551	0.914
Mean	***2.693***	***2.601***	***2.434***
Neurosciences			
Annu Rev Neurosci	26.676	30.167	28.533
Nat Rev Neurosci	-	27.007	23.054
Behav Brain Sci	14.250	10.625	14.964
Nat Neurosci	12.636	15.141	14.805
Neuron	15.081	14.109	13.894
Trends Neurosci	17.417	12.631	13.494
Mol Psychiatr	8.927	5.539	11.804
Front Neuroendocrin	-	8.870	11.526
Prog Neurobiol	9.933	12.327	11.304
Trends Cogn Sci	-	7.528	9.374
Mean	***14.989***	***14.394***	***15.275***
Obstetrics and Gynecology			
Hum Reprod Update	2.887	3.731	6.793
Obstet Gynecol	2.091	2.957	3.813
Hum Reprod	2.997	3.125	3.769
Obstet Gynecol Surv	-	1.773	3.329
Fertil Steril	2.854	3.483	3.277
Reprod Biomed Online	-	-	3.206
Menopause	2.273	3.319	3.170
Semin Reprod Med	-	1.575	3.000
Placenta	2.587	2.706	2.969
Am J Obstet Gynecol	2.519	2.518	2.805
Mean	***2.601***	***2.799***	***3.613***
Oncology			
Ca-Cancer J Clin	24.674	33.056	63.342
Nat Rev Cancer	-	33.954	31.583
Cancer Cell	-	18.913	24.077
J Natl Cancer I	14.159	13.844	15.271
J Clin Oncol	8.773	10.864	13.598
Adv Cancer Res	21.680	7.938	10.682
Lancet Oncol	-	7.411	10.119

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
BBA-Rev Cancer	-	8.395	9.156
Stem Cells	2.989	5.802	7.924
Cancer Res	8.460	8.649	7.656
Mean	***13.456***	***14.883***	***19.341***
Ophthalmology			
Prog Retin Eye Res	4.680	6.811	9.039
Ophthalmology	3.040	3.162	4.031
Invest Ophth Vis Sci	4.373	4.148	3.766
J Vision	-	-	3.753
Surv Ophthalmol	2.562	3.096	3.451
Arch Ophthalmol-Chic	2.158	3.203	3.206
Exp Eye Res	2.014	2.611	2.776
Brit J Ophthalmol	1.948	2.099	2.524
Am J Ophthalmol	1.941	2.258	2.468
Mol Vis	-	2.777	2.377
Mean	***2.840***	***3.352***	***3.739***
Orthopedics			
Osteoarthr Cartilage	2.080	2.964	4.017
J Orthop Res	-	2.167	2.784
Orthop Clin N Am	0.874	0.907	2.500
J Bone Joint Surg Am	2.222	1.921	2.444
Spine	1.843	2.676	2.351
Clin Orthop Relat R	1.182	1.357	2.161
Gait Posture	0.955	1.585	1.976
Eur Spine J	-	1.527	1.824
J Arthroplasty	0.978	0.922	1.806
J Am Acad Orthop Sur	-	-	1.792
Mean	***1.448***	***1.781***	***2.336***
Otorhinolaryngology			
Jaro-J Assoc Res Oto	-	2.086	2.522
Head Neck-J Sci Spec	1.917	1.805	1.961
Ear Hearing	1.506	1.450	1.858
Arch Otolaryngol	1.527	1.242	1.816
Audiol Neuro-Otol	2.390	1.765	1.758
Laryngoscope	1.457	1.449	1.736
Hearing Res	1.753	1.502	1.584
Otol Neurotol	-	1.073	1.339
Otolaryng Head Neck	0.977	1.051	1.338
Am J Rhinol	1.021	1.055	1.220
Mean	***1.569***	***1.448***	***1.713***
Pathology			
Am J Pathol	6.971	6.946	5.917
J Pathol	4.137	4.933	5.759
Brain Pathol	6.435	3.838	5.274
Springer Semin Immun	2.176	0.918	4.754
Lab Invest	4.165	4.418	4.453
J Neuropath Exp Neur	5.565	5.005	4.371
Am J Surg Pathol	4.269	4.535	4.144
Modern Pathol	3.241	3.323	3.753
Histopathology	2.554	2.952	3.216
Int J Immunopath Ph	1.174	3.927	3.213
Mean	***4.069***	***4.080***	***4.485***

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
Pediatrics			
Pediatrics	3.742	3.781	5.012
J Am Acad Child Psy	3.175	3.779	4.767
J Pediatr	3.467	2.913	3.991
Arch Pediat Adol Med	1.701	2.190	3.565
Pediatr Infect Dis J	2.190	2.262	3.215
Pediatr Allergy Immu	1.635	1.573	2.849
J Adolescent Health	1.415	1.674	2.710
Ment Retard Dev D R	0.811	3.479	2.671
Pediatr Res	2.794	3.064	2.619
J Child Adol Psichop	1.982	2.487	2.486
Mean	***2.291***	***2.720***	***3.389***
Peripheral Vascular Disease			
Circulation	10.893	11.164	10.940
Circ Res	9.193	10.117	9.854
Artherioscl Throm Vas	5.111	6.791	6.883
Hypertension	5.311	5.630	6.007
Atherosclerosis supp	-	4.457	5.875
Curr Opin Lipidol	5.661	6.966	5.689
Stroke	6.008	5.233	5.391
J Thromb Haemost	-	-	5.138
Curr Opin Hephrol HY	2.544	3.976	4.137
J Hypertens	3.640	3.572	4.021
Mean	***6.045***	***6.434***	***6.394***
Pharmacology and Pharmacy			
Annu Rev Pharmacol	19.289	21.786	22.808
Nat Rev Drug Discov	-	17.732	20.970
Pharmacol Rev	25.381	27.067	16.854
Trends Pharmacol Sci	10.377	13.965	10.400
Pharmacol Therapeut	6.487	7.397	8.657
Clin Pharmacol Ther	5.275	6.141	8.066
Adv Drug Deliver Rev	2.406	6.588	7.977
Pharmacogenetics	4.465	5.851	7.221
Med Res Rev	3.417	7.788	7.218
Drug Discov Today	4.105	4.943	7.152
Mean	***9.022***	***11.926***	***11.732***
Physiology			
Physiol Rev	27.677	36.831	31.441
Annu Rev Physiol	18.848	18.591	15.356
Physiology	-	-	6.268
Rev Physiol Bioch P	5.389	6.333	5.625
News Physiol Sci	2.060	3.682	5.241
J Gen Physiol	6.082	5.120	4.962
Pflug Arch Eur J Phy	-	-	4.807
J Biol Rythm	2.867	4.061	4.633
J Physiol-London	4.455	4.352	4.407
Am J Physiol-Cell Ph	4.086	4.103	4.334
Mean	***8.933***	***10.384***	***8.707***
Psychiatry			
Arch Gen Psychiat	11.778	10.519	13.936
Mol Psychiatr	8.927	5.539	11.804
Am J Psychiat	6.577	7.157	8.250

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
Biol Psychiat	4.269	6.039	7.154
Neuropsychopharmacol	4.579	5.201	5.889
J Clin Psychiat	4.454	4.978	5.533
Brit J Psychiat	4.827	4.421	5.436
Int J Neuropsychoph	1.323	4.000	5.184
J Am Acad Child Psy	3.175	3.779	4.767
J Clin Psychopharm	5.052	4.432	4.561
Mean	***5.496***	***5.607***	***7.251***
Rehabilitation			
Neurorehab Neural Re	0.190	-	2.403
J Rehabil Med	-	1.068	2.168
Manual Ther	-	1.189	1.931
Support Care Cancer	1.174	1.367	1.905
IEEE T Neur Sys Reh	-	1.270	1.842
Arch Phys Med Rehab	1.409	1.350	1.826
Phys Med Rehab Kuror	0.160	0.485	1.746
J Burn Care Rehabil	0.810	1.042	1.744
J Electromyogr Kines	1.146	1.352	1.725
J Orthop Sport Phys	1.424	1.036	1.525
Mean	***0.902***	***1.129***	***1.882***
Respiratory System			
Am J Resp Crit Care	5.443	8.876	9.091
Thorax	3.979	4.188	6.064
Eur Respir J	2.590	2.999	5.076
Am J Resp Cell Mol	4.353	4.015	4.593
Am J Physiol-Lung C	3.303	3.735	4.250
Chest	2.451	3.264	3.924
J Thorac Cardiov Sur	3.057	3.319	3.560
Lung Cancer	-	-	3.554
Tuberculosis	-	1.594	3.425
J Heart Lung Transpl	2.526	2.843	2.830
Mean	***3.463***	***3.870***	***4.637***
Rheumatology			
Arth Rheum/Ar C Res	-	7.190	7.751
Ann Rheum Dis	2.444	3.827	5.767
Curr Opin Rheumatol	-	3.150	4.805
Rheumatology	2.537	3.760	4.052
Osteoarthr Cartilage	2.080	2.964	4.017
Arthritis Res Ther	-	5.036	3.801
Semin Arthritis Rheu	3.066	2.598	3.440
J Rheumatol	2.910	2.674	2.940
Rheum Dis Clin N Am	2.257	2.776	2.568
Lupus	2.514	1.808	2.366
Mean	***2.544***	***3.578***	***4.151***
Spectroscopy			
Mass Spectrom Rev	7.600	7.364	10.947
Prog Nucl Mag Res Sp	5.062	5.971	6.417
Appl Spectrosc Rev	0.500	1.000	3.846
J Anal Atom Sprectrom	3.488	3.200	3.630
NMR Biomed	1.914	3.333	3.626
J Am Soc Mass Spectr	3.040	3.321	3.307
Spectrochim Acta B	2.608	2.361	3.092

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
J Mass Spectrom	2.638	2.875	2.945
Rapid Commun Mass Sp	2.184	2.789	2.680
Int J Mass Spectrom	1.923	2.361	2.337
Mean	***3.096***	***3.458***	***4.283***
Surgery			
Ann Surg	5.987	5.937	7.678
Am J Transplant	-	5.678	6.843
Liver Transplant	2.130	4.242	4.629
Am J Surg Pathol	4.269	4.535	4.144
Brit J Surg	2.935	3.772	4.092
Transplantation	4.035	3.608	3.972
Obes Surg	1.464	2.421	3.723
J Neurol Neurosur Ps	2.846	3.035	3.630
Endoscopy	1.817	3.227	3.605
J Thorac Cardiov Sur	3.057	3.319	3.560
Mean	***3.171***	***3.977***	***4.588***
Toxicology			
Annu Rev Pharmacol	19.289	21.786	22.808
Mutat Res-Rev Mutat	4.129	5.783	7.579
DNA Repair	-	3.277	5.868
Toxicol Appl Pharm	2.730	2.851	4.722
Drugs	3.966	4.611	4.472
Mutat Res-Fund Mol M	2.148	3.433	4.111
Crit Rev Toxicol	6.360	2.471	3.707
Drug Safety	2.763	2.971	3.673
Toxicol Sci	2.361	3.067	3.598
Chem Res Toxicol	3.187	3.332	3.162
Mean	***5.215***	***5.358***	***6.370***
Transplantation			
Am J Transplant	-	5.678	6.843
Liver Transplant	2.130	4.242	4.629
Transplantation	4.035	3.608	3.972
Cell Transplant	2.959	2.327	3.482
Biol Blood Marrow Tr	-	2.880	3.458
Nephrol Dial Transpl	2.056	2.607	3.154
Stem Cells Dev	-	-	3.076
J Heart Lung Transpl	2.526	2.843	2.830
Bone Marrow Transpl	2.396	2.172	2.621
Transpl Immunol	1.453	1.075	2.297
Mean	***2.508***	***3.048***	***3.636***
Tropical Medicine			
Malaria J	-	-	2.748
Trop Med Int Health	1.350	2.156	2.595
Am J Trop Med Hyg	1.765	2.105	2.546
Acta Trop	0.799	1.336	2.211
T Roy Soc Trop Med H	1.485	2.114	2.030
Mem I Oswaldo Cruz	0.542	0.688	1.208
Ann Trop Med Parasit	0.988	1.010	1.191
Ann Trop Paediatr	0.413	0.704	0.934
Leprosy Rev	1.343	0.907	0.847
J Trop Pediatrics	0.447	0.514	0.592
Mean	***1.015***	***1.282***	***1.690***

(continued)

(continued)

Sector / Journal	IF 2000	IF 2003	IF 2006
Urology and Nephrology			
J Am Soc Nephrol	5.745	7.499	7.371
Eur Urol	2.058	2.247	4.850
Kidney Int	4.371	5.302	4.773
J Sex Med	-	-	4.676
Am J Phisiol-Renal	4.129	4.344	4.199
Curr Opin Nephrol Hy	2.544	3.976	4.137
Am J Kidney Dis	3.646	3.897	4.072
J Urology	2.896	3.297	3.956
Prostate	3.754	3.278	3.724
Eur Urol Suppl	-	-	3.174
Mean	***3.643***	***4.230***	***4.493***
Overall mean	***4.913***	***5.538***	***6.158***

Decreasing order according to the 2006 IF.
Source: Journal Citation Reports™ – Science edition, published by Thomson Scientific (with permission).

10.3. Do We Always Need Institutional Review Board Approval and Informed Consent?

The answer is: Yes

If we consider prospective studies on humans, the answer to this question is: Yes, we do need them. We cannot perform here a deep analysis of the ethical, deontological, and regulatory problems related to medical research on humans. Instead we refer the reader to the *Helsinki Declaration*, approved in 1964 by the *World Medical Association*, and later emended in 1975 (Tokyo), 1983 (Venice), 1989 (Hong Kong), 1996 (Somerset West, South Africa) and most recently in 2000 (Edinburgh). Clarifications on particular paragraphs were adopted in 2002 (Washington) and in 2004 (Tokyo) [World Medical Association, 2004]. Notice that the most recent version confirms and emphasizes the importance of informed consent by the patients enrolled in a study.

Helsinki Declaration

What are the general rules which need to be respected?

All prospective studies need preliminary approval by the Ethics Committee or Institutional Review Board and informed consent must be obtained from all patients for them to be enrolled in the study and for their data to be managed for scientific purposes. One may think that this rule holds only for those studies where patients are randomized to one of two or more diagnostic modalities or for those studies where the patients undergo a new, "experimental", imaging technique they would not undergo according to a standard of care diagnostic algorithm. For instance, you might imagine that a prospective study comparing Doppler ultrasound and MR angiography in patients with suspected carotid stenosis performed as routine diagnostic examinations could be done and published without any Institutional Review Board approval. This is not the case, and not only because we might be lacking informed consent to use patient data. In fact, the prevalent current opinion is that Institutional Review Board approval (and not only informed consent) is an absolute requirement for all prospective studies.

Moreover, this rule is also valid for retrospective studies, where we gather, manage, and analyze the data after the diagnostic events, even many years

For every study on humans we need the IRB approval

after. In situations like these, it is just the Institutional Review Board approval which makes the study publication possible especially when informed consent by the patients is not available: some of them may have moved to another country or city or may have died (as is the case in oncology).

In our experience, we request the approval for a retrospective study from the local Ethics Committee with a short document including the nomination of the person responsible for personal data. Within two or three weeks, the study leader meets the Ethics Committee for an oral presentation of the study. The approval is almost always immediate.

An indirect demonstration of the necessary Institutional Review Board approval of both prospective and retrospective studies on humans is the request for a statement outlining that approval has been obtained for the submission of an abstract of a scientific papers or a poster to the Annual Meeting of RSNA. More importantly, this statement is today an absolute must for the acceptance for publication of any prospective or retrospective study on humans in any well respected journal. Many journals ask the reviewers to make a particular check within the text for evidence of Institutional Review Board approval, typically placed at the beginning of the section regarding materials and methods. Some radiologic journals requires this declaration to also be placed in the abstract.

The need of IRB approval compels higher quality research

This development should be seen as a positive trend, regardless of ethical concerns. In fact, the need of Institutional Review Board approval prompts a preliminary analysis of the quality of the research project, careful reading of the literature, precise definition of the study protocol, and the exchange of views with all the members of the Institutional Review Boards or Ethics Committee. This book could be useful to radiologists for this exchange of opinions, but we strongly suggest studies be designed in close cooperation with statisticians, especially studies with patient randomization and sample size calculation.

10.4. Title, Running Title and Title Page

The title is important

The title of the paper is important. During the 1980s, the ratio between the number of people who read the whole text of a paper and those who read only the title was estimated to be about 1:500 [Kerkut, 1983]. Today access to the medical literature via the internet may have further reduced this ratio.

A title should stimulate the reader to read at least the abstract. Titles are mostly a short description of the subject matter contained in the article. An alternative which is still little utilized in radiologic journals is the declarative title.

Neutral or declarative titles

This title makes a brief statement on the results of the study [Gustavii, 2003]. A paper clearly demonstrating a higher overall accuracy of CT compared to ultrasound in the diagnosis of liver metastases in 135 patients affected with colorectal cancer can be modestly entitled: *Ultrasound and computed tomography in the diagnosis of liver metastases*. The assertive alternative is: *Computed tomography is more accurate than ultrasound in the diagnosis of liver metastases*. The second option, communicating the main study result, has a stronger impact. However, bear in mind that some important journals (e.g. *JAMA* and *New Engl J Med*) require descriptive and not assertive titles.

Interrogative titles

We can also opt for an interrogative title. For our example: *Are there differences in overall accuracy between ultrasound and computed tomography in the*

diagnosis of liver metastases? Is ultrasound less accurate than computed tomography for the diagnosis of liver metastases? or *Is computed tomography more accurate than ultrasound for the diagnosis of liver metastases?* For an original article we prefer a descriptive neutral title or an assertive title containing the answer to the question. An interrogative title can be highly suitable for a narrative review which considers different answers to the question.

Subtitle

Some journals (e.g. *Invest Radiol*) accept that the title is followed by a clarifying subtitle. For our example: *Computed tomography is more accurate than ultrasound in the diagnosis of liver metastases. A prospective study on 135 consecutive patients affected with colorectal cancer.* This option can also be adopted with a colon placed between the first and the second sentence. In this way, we have a brief summary of the content of two sections of the paper (*results: materials and methods*).

In the title, avoid abbreviations or acronyms (i.e. words formed by the initial letters of different words), unless – in radiologic journals – they are well recognized such as CT, MR, or US, etc. Note that in the title of a congress abstract we can use new acronyms (introducing them in brackets after the first full form) in order to save space in abstract text. Eliminate these acronyms from the title of the full paper to be submitted to a journal.

Running title

Many journals require a *running title*, namely a short title (frequently no more than 50 characters, blank spaces included; acronyms generally permitted) to be placed at the upper margins of the journal pages. For example: *Liver metastases: CT and US accuracy* (35 characters).

Full title page

Manuscript editing (always done using *word processing* software) requires a *full title page* containing: the title of the paper; first and family names and affiliations of the authors; type of article (e.g. *original article* or *original research*) and the complete address (including telephone number, fax number and e-mail address) of the corresponding author who is submitting the paper and to whom the readers could refer for any discussion or request of reprints. This full title page should be a separate file not to be submitted to the reviewers, permitting a blind review process. For this reason, a *blind title page* (with the manuscript title only) is to be placed before the beginning of the main body of the manuscript.

Blind title page

10.5. Four-section Scheme, Section Size and Editing Sequence

Main body

An *original article* always follows a four-section scheme including: *Introduction*; *Materials and Methods*; *Results*; and *Discussion*. The set of these four sections is named *main body.* Some nominal variations are *Background* instead of introduction; *Methods* or *Subjects and Methods* instead of materials and methods, etc. but the content of the sections is the same. Conclusions, which in the abstract substitute for the entire discussion, is usually the final part of discussion and not a separate section.

Section size

Beginners should take into consideration a general rule for the size of each of the four sections and carefully read the *instructions for authors* of the journal to observe possible limits to the whole size of the manuscript. Using a normal word processor, A4 format page with 2-cm margins, 12-point *Times New Roman* type characters, and double spacing, the maximum size of the sections of a typical manuscript should be as follows:

1. *Introduction*: 1-2 pages;
2. *Materials and Methods*: 3 pages;
3. *Results*: 3 pages;
4. *Discussion*: 3-4 pages.

Manuscript size

Note that the total size of the four sections (about 10 pages) is usually half the size of the whole manuscript including also blind title page, abstract, references, tables (each of them placed in a new page) and figure legends. Some components (usually tables and figures), may be required as separate files. Remember that some journals put a limit on the main body given by the sum of the four sections (e.g. 3,000 words for *Radiology*, 4,500 word for *AJR Am J Roentgenol*).

Lastly, you do not need to imitate the typographic style of the journal (printed characters, paragraph indentation, etc.). You can refer to the *Vancouver Requirements* (www.icmje.org) or edit the text according to the rules listed in this chapter.

10.6. «Introduction»: Why did you do It?

Do not begin with the abstract

In the practical writing and editing of a scientific manuscript we should not follow the logical sequence of the printed paper. In particular, to begin with the abstract is not a good idea, not even when you have the abstract accepted at a congress. Sometimes this abstract is much larger than that permitted by the journal. More importantly, the abstract must effectively represent quality of methods as well as originality of results, and respect narrow mandatory limits (see Section 10.11). This ambitious aim can be reached only when the entire manuscript has been completed.

Can you begin with the introduction, as formal logic may suggest? The initial part of the text of the research project approved by the Institutional Review Board is a good starting point, even though it is commonly too long to be placed without shortening and modification as an introduction. We suggest (especially for beginners) writing or refining the introduction after editing the materials and methods and results sections. This choice could also be useful considering the partial overlap between the topics of the introduction and those of the discussion, especially regarding the reference to previously published papers.

Why did you do it?

The introduction should answer the question: Why did you do it? We stated above that it should be no longer than one or two pages, with one page corresponding to about 300 words, but 400-500 words is common enough. The easiest scheme for an introduction answers two simple questions: What is the problem? What did you do to solve it? [Gustavii, 2003]. Another scheme is composed of three blocks, each to be ended with a full stop and followed by a new paragraph:

Three-block Introduction

1. general background (e.g. epidemiology of the disease);
2. particular background (e.g. diagnostic performance of standard of care imaging modalities);
3. purpose of the paper (e.g. to evaluate sensitivity and specificity of a new imaging modality).

Four-block Introduction

A four-block scheme could be:

1. a problem exists in diagnosing disease X with technique Y;
2. previous efforts have been made by other authors to solve the problem;
3. the results obtained with a new approach (new Y) were obtained in a different clinical field;
4. aim: to evaluate the performance of new Y for the diagnosis of X.

Beginning the Introduction

A useful word of advice is to begin with a short statement summarizing the problem or the context of the problem. To come back to the example of liver metastases, an initial statement could be: *The knowledge of number and location of liver metastases is crucial for treatment planning in patients with colorectal cancer.* You must avoid the temptation of a long introduction, which is especially dangerous when you have wide knowledge of the matter under investigation. Bear in mind this simple rule: move to the discussion what cannot be included in the introduction.

Ending the Introduction

At any rate, the introduction must end with a paragraph stating the purpose of the study (you may take it from the purpose of the congress abstract).

10.7. «Materials and Methods»: What did you do and how did you do it?

Supply all information useful to reproduce your study

The beginner who is going to write the materials and methods (from now on simply "methods") of a manuscript should have a broad understanding of the basic aim of this section. Many residents and young radiologists are astonished by the very high level of detail required for reporting methods in radiologic journals. A summary description of what you have done and how you have done it is absolutely inadequate. You must give all the information which enables other colleagues to repeat your study on a similar sample of patients and, therefore, to confirm or deny your results.

One swallow doesn't make a summer

Here we should come back to a general principle already discussed in Chapter 3 which is encapsulated in the proverb: *One swallow doesn't make a summer*. A new promising result reported by a group of authors – for instance the high accuracy of a new imaging modality for the diagnosis of an important disease – needs multiple confirmations by other groups of authors before it can be declared clearly demonstrated. This mechanism of medical science implies that researchers must know the *exact* experimental conditions of the study they would like to reproduce. Only a partial exception to this rule is possible when your methods have been previously described in detail by you or other authors. In this case, you can use the phrase "as already described" followed by the number of the corresponding reference. The exception is partial because it is common for the editor or reviewers to ask you to at least summarize the already published methods while keeping the references.

Methods subsections and subheadings

We advise subdividing methods into subsections, each with subheadings you can usually freely choose in relation to their content. The most commonly used are the following: *study design*; *study population*; *imaging protocol*; *imaging analysis*; *standard of reference* (not always corresponding to *pathologic examination*); *radiologic-pathologic correlation*; *statistical analysis*. But many vari-

ations are possible. This subdivision is suitable even when each subsection is composed of only a few lines of text. Reading the text will be easier and the reader will more rapidly find a particular methodologic aspect.

Study design

For *study design*, see Chapter 8.

Study population

In the *study population* subsection you should provide sufficient information to make clear whether your results can be applied to another population of patients. You should report not only age and sex but also the clinical status and all the inclusion and exclusion criteria (to allow for a definition of the pretest disease probability – see Section 1.4) as well as the consecutive or nonconsecutive modality of enrollment. Here you should state having obtained Institutional Review Board approval and informed consent from the enrolled patients. Moreover, if the patients were randomized, you should explain the modality of randomization and the level of blindness of the study (see Chapter 8). The study time period should be declared reporting month and year of the first and last enrollments.

Flow diagram

The course from enrollment to results can be usefully summarized in a flow diagram, frequently required by the journals. An example is reported in Figure 10.1. A detailed reconstruction is particularly needed for large randomized trials but should be more extensively applied, including the number of patients screened, excluded (and the reasons for excluding them), eligible, refusing consent, randomized to each arm of the study, retired (dropout), and concluding the study for each arm. Notice that the number of excluded patients (how many for a contraindication to the examination under investigation?) can also be of great value for radiologic prospective nonrandomized studies. It is important to define the possibility of applying the results of these studies to clinical practice. This number is sometimes not provided in radiologic studies, as happens when the authors list the exclusion criteria but do not report the number

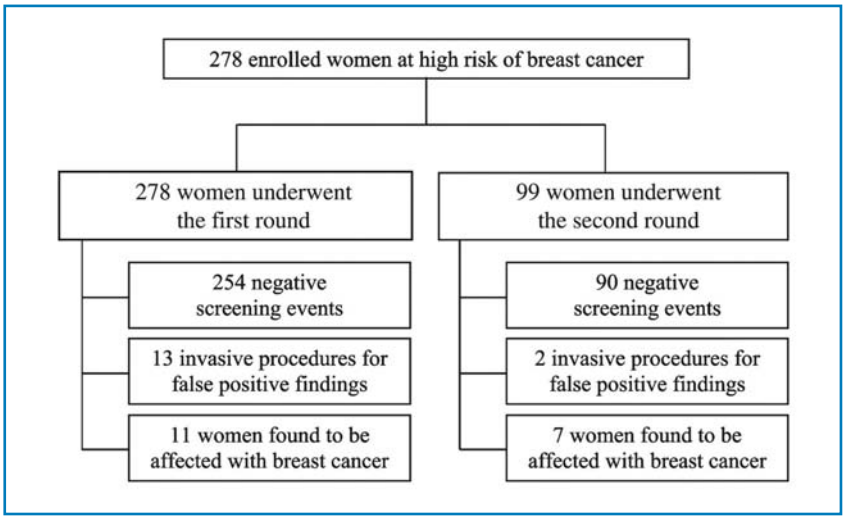


Figure 10.1. Flow diagram of a nonrandomized study. This scheme provides a synopsis of the distribution of a sample of 278 women at high genetic-familial risk of breast cancer at the first and second round of a multimodality surveillance program and the general results [Sardanelli et al (2007) Radiology 242:698-715, with permission of the copyright owner (RNSA)].

of screened patients as well as the number of excluded patients with the reasons for exclusion, thus making the difference between screened and eligible patients impossible to calculate. A theoretical example of this more complete flow diagram can be seen in Figure 10.2.

The first two subsections of the methods can be merged into a single subsection entitled *study design and population*, which would also include the description of the control group where present. Alternatively, a special subsection entitled *control group* can be added.

Imaging protocol

You should accurately describe how you performed the imaging modalities in the subsection entitled *imaging protocol*, including: brand name of the manufacturers of the equipment (with city and country of the registered office); model of the equipments (also for power contrast injectors); release of software; and all technical parameters. If these data are repetitive (as they may be for MR sequences), you can create one or more tables. Here (or in the subsection entitled *imaging analysis*) you should declare specific training and years of experience of the radiologists who performed and interpreted the examinations. If a study includes multiple imaging modalities, we advise creating multiple subsections, one for each imaging modality.

Imaging analysis

Imaging analysis has an importance similar to that of imaging protocol. Here you should describe who interpreted the images (training and specific experience) and how they were interpreted, whether using hardcopy or softcopy

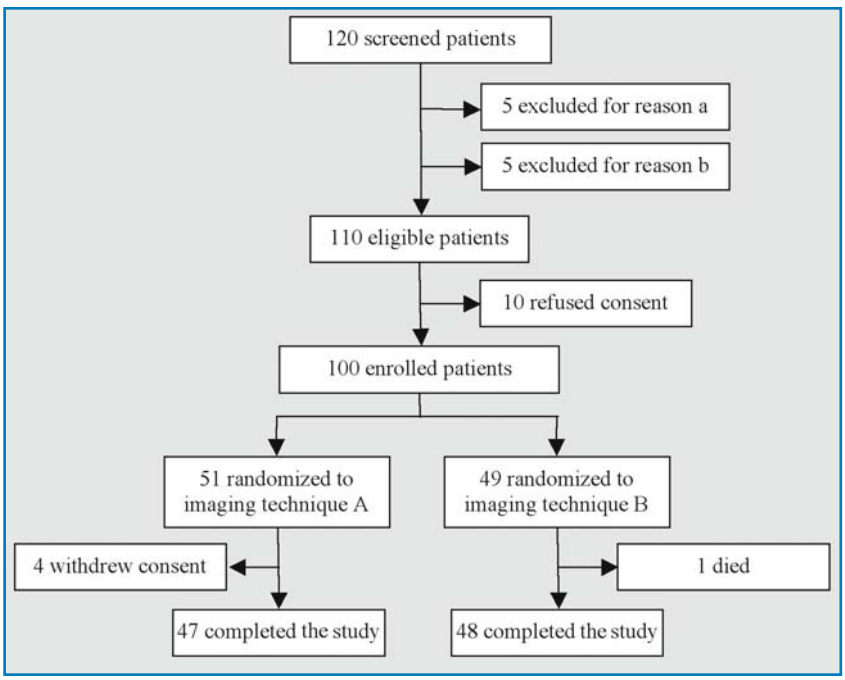


Figure 10.2. Flow diagram of a randomized study. This scheme shows a progressive reduction in the number of patients: from 120 screened patients to 47 patients who completed the study with imaging technique A and 48 who completed the study with imaging technique B.

methods, with which windowing on the video-display, with or without knowledge of clinical information or results of previous examinations, hopefully blinded to the reference standard. If the observer(s) have done multiple readings (as is the case for studies on intraobserver reproducibility – see Chapter 7), you should state the time interval between the readings and which procedures were adopted to avoid that the reader had memory of the first reading at the moment of the second reading (e.g. by means of randomized reading order). You should declare how an examination was defined positive or negative, namely the diagnostic criteria for the use of categorical or ordinal variables (see Chapter 2), with possible references to the use of these methods in previously published papers. If you measured continuous variables (e.g. lesion size, CT density, signal-to-noise or contrast-to-noise ratio, etc.), you should report the procedure used and the way of calculating variables and indices. If you used special software for imaging analysis, you should declare details as already stated for the equipment. If special signs or imaging features or particular technical procedures are used (e.g. innovative software for image segmentation), one or two figures can be associated to methods for a better understanding of the procedure.

Reference standard

The classic *standard of reference* is histopathology. However, this may be impossible, in part for ethical reasons. Negative examinations, especially for asymptomatic subjects in screening programs, are compared to clinical and imaging follow-up. In other circumstances, the definitive diagnosis is obtained using the findings of another imaging modality considered a *standard of care* for the investigated disease or using a combination of multiple evaluations (*final assessment*). You must explain all of this in detail. If the paper includes a pathologic reference standard, in this subsection (which can be named *pathologic examination*) you should report the histologic techniques and diagnostic criteria used, with reference to previously published papers (here ask the pathologist for advice, even if she/he is not an author).

Final assessment

Radiologic-pathologic correlation

Notice that in the presence of a histopathologic reference standard, the topographic correlation between pathologic and imaging findings can be problematic, especially in cases of multiple lesions in the same organ or segment. In these situations (for instance, liver metastases or multifocal/multicentric breast cancers), methods for *radiologic-pathologic correlation* should be clearly explained.

Statistical analysis

Statistical analysis is almost always the final subsection of the methods. Here you should explain how true and false positives and negatives were defined. Moreover, you should state which statistical tests were used, which α error was adopted (usually, $p \leq 0.05$ – see Chapter 3) and, hopefully for prospective studies, the study power (usually 0.80-0.90 – see Chapters 3 and 8). The power calculation should be considered mandatory at least in the case of nonsignificant results. The choice of statistical tests, especially if they are not usual, should be justified with reference to previously published papers. In particular, the use of parametric tests should be justified by means of a preliminary demonstration of the existence of the necessary conditions (normal data distribution in the sample or assumption of normal distribution in the population, etc. – see Chapters 2 and 4). Lastly, you should state which statistical software package was used (specify the release number, city and country of the registered office of the manufacturer).

10.8. «Results»: What did you Find?

Perfect correspondence between Methods and Results

The first rule to illustrate *what you found* is the following: *all the results must have the description in the methods of the way you found them*. Conversely, all the described methods must have in the results the description of the findings obtained using them. The second rule is that *results must be presented in a neutral manner, without any adjective or comment minimizing or emphasizing their meaning or possibility to be clinically applied.* A subdivision in subsections with specific subtitles is also welcome in the results section.

Tables and graphs

The use of tables and/or graphs is highly recommended. Do not report data in the text which is already provided in tables and graphs. For text continuity, you may provide some summary data drawn from the tables. In practice, the results are sometimes limited to a few lines referring to tables and graphs.

The creation of tables and graphs is no trivial task, both for conceptual and technical issues. Try using paper and pen first of all to manually design an effective scheme. Then go to the computer and use the proper function of the word processor (not free text with tabulations and blank spaces) to avoid problems in the pdf file submission and to facilitate the future editorial work by the journal office. Tables should make reading the data easy and they should be organized according to the journal style (as close as possible). You can use tables recently published in the same journals as a model. If you have a large amount of data, use two, three, or more tables. Remember that in any table, *percentages* must be accompanied by their engendering ratio.

Percentages

Suitable and well designed graphs can visually represent data and results in a more effective way than any free text or table. Regardless of the software you use (Excel®, or other statistical packages), you can place the graphs in the main file of the manuscript or save them as images using other software (e.g. Adobe Photoshop®). We advise using a tiff digital format (or other permitted software, according to the instructions for authors) and saving the graphs with high spatial resolution, up to 1200 dots per inch (dpi), to avoid them being refused by the editorial office and/or appearing blurred in the printed paper. This is also valid for the flow diagrams described in Section 10.7.

Figures

All *figures* containing radiologic images (commonly one figure is composed of multiple panels) are typically placed in the results. They must be of high quality, limited in number, representative of the principal message of the paper, and conveniently magnified and cropped to allow the findings to be clearly recognized. Arrows and other graphical signs should be used to indicate the findings (also those immediately evident to the authors), in accordance with the journal style.

Legends

All *tables* must be numbered and need a title and, frequently, also notes and explanations of symbols, in accordance with the journal style. All the figures (line art designs, graphs, radiologic images) must be numbered and need a legend which can never be reduced to a reference to the text. The golden rule for tables (with titles and notes) and figures (with legends) is: *they must be understandable by a reader who observes them without having read the text of the paper, that is they must provide enough information to bring out the message by themselves*. This is the reason for repeating information on the

patients sample in a table title and not to use acronyms in tables and figure legends unless they are explained in the same table or in the figure legend (see Section 10.12).

Each table or figure, being a logical extension of the text of the manuscript, must be quoted at least in one point of the text of the manuscript. Typically, the references to a table or figure after the first one are indicated in brackets as "(see Figure X)".

All the significances or non significances obtained with statistical tests should be placed in the results. Remember the need to provide the *p* values at least up to the third decimal place, whether they are significant or nonsignificant (see Chapter 3).

10.9. «Discussion»: What is the Meaning of your Findings?

A creative work

The discussion is probably the most creative part of the whole paper, apart from the original idea (engendering the experimental hypothesis) and the design of the study. As a consequence, it is also the most freely structured. Beginners need some help from more experienced colleagues. A general rule is: *You must discuss your results, not show your general culture on the matter you investigated. The discussion is aimed at interpreting and commenting on your results.*

Two ways for beginning the Discussion

Two possible ways to begin the discussion are commonly used. The first approach is to summarize your results. In the liver metastases example: *The current study demonstrated that contrast-enhanced CT is more accurate than ultrasound in diagnosing liver metastases in patients with colorectal cancer.* The second approach is to come back to arguments already presented in the introduction, maybe from a different viewpoint: *The incidence of colorectal cancer is increasing in all developed countries, shown by recent epidemiologic reports… [references].*

Discuss your results, point by point

At any rate, after a short "introduction to the discussion", you must comment on your work. If there are aspects of your methods to be discussed, you may do so. But *the core of this section is the discussion, point by point, of your results.* A practical word of advice: print your results section (including tables) on paper, read it carefully and then edit a comment of each point, comparing your work with already published papers on a similar topic (with a reference for each study you quote). Try to explain the reasons for the differences between your results and those obtained by other authors, whether they are better or worse than your own. If there are no important differences, state that you have confirmed the results already obtained by others. If some of your results are trivial or obvious, comment on them by stating "as expected". In the discussion section you can finally comment on the quantities (the numbers) of your results in a qualitative way. Try to indicate which clinical implications, which effects on patient care your study could have as well as which aspects deserve future investigations. You can highlight the original aspects of your work, but be cautious in claiming to be the first to demonstrate a research hypothesis (see Section 10.13).

Study limitations

Do not forget to list the *limitations* of your study, and try to detect possible sources of bias (see Chapter 9), which may be unavoidable for various reasons

(ethical, logistic, organizational, others…) to state in your discussion. The readers must be warned against applying your results to a different clinical or epidemiological context.

Lastly, most papers end the discussion with a statement summarizing the major findings of the study, which serves as a conclusion.

10.10. «References»

Your business card to the reviewers

References are the business card of your manuscript to the reviewers. Their choice and the accuracy in their editing according to the journal style are a strong indicator of the quality of the study. Inaccurate references create a negative bias in the reviewers' evaluation. Some practical rules include:

1. verify possible explicit limits in the references number stated in the instructions for authors;
2. choose the most important and recent papers of a group of authors who have published many papers on the same topic;
3. always use references accessible on MEDLINE/PubMed, with the only exception of books, chapters in book, and congress abstract for particular cases[4];
4. do not use *second hand references*, namely do not copy references from other studies, to avoid spreading errors, especially if the sources are old papers;
5. extract from PubMed all the references using a "copy and paste" technique (after having activated the "send-to-text" option), to avoid errors, which otherwise very commonly occur;
6. respect all the rules of the journal for the reference format, also those for the citation of books and book chapters as well as for online journals and web sites (for which the access date is requested); unpublished material or personal communications should be referenced in the free text, in brackets;
7. if you can, use dedicated software (e.g. EndNote ®) which gives you an automated formatting for the journal style (but always visually check the result).

Respecting the rules of the journal for the reference format is important not only because some editors ask the reviewer to do this check, but also because:

1. even though the authors are responsible for the accuracy of the references, many journals make automated controls and ask for the correction of all the irregular references;
2. a "wrong" reference format can reveal that the manuscript has been previously submitted to another journal which rejected it.

Do not reveal the previous refusal

In Table 10.5 we show how the same reference should be reported for five radiologic journals.

[4] Particular cases of relevant results which at the moment of the manuscript submission have been published only as congress abstract. In other cases, prefer always the reference to a journal than that to a congress abstract.

Table 10.5. Format for references. The same paper, an original article, is reported according to the instructions of five radiologic journals

Radiology	Sardanelli F, Iozzelli A, Losacco C, Murialdo A, Filippi M. Three subsequent single doses of Gd-chelate in brain MR of multiple sclerosis. AJNR Am Journal Neuroradiol 2003;24:658-662.
Invest Radiol	Sardanelli F, Iozzelli A, Losacco C, et al. Three subsequent single doses of Gd-chelate in brain MR of multiple sclerosis. *AJNR Am Journal Neuroradiol*. 2003;24:658-662.
J Magn Reson Imaging	Sardanelli F, Iozzelli A, Losacco C, Murialdo A, Filippi M. Three subsequent single doses of Gd-chelate in brain MR of multiple sclerosis. AJNR Am Journal Neuroradiol 2003;24:658-662.
Eur Radiol	Sardanelli F, Iozzelli A, Losacco C, Murialdo A, Filippi M (2003) Three subsequent single doses of Gd-chelate in brain MR of multiple sclerosis. AJNR Am Journal Neuroradiol 24:658-662
AJR Am J Roentgenol	Sardanelli F, Iozzelli A, Losacco C, Murialdo A, Filippi M. Three subsequent single doses of Gd-chelate in brain MR of multiple sclerosis. *AJNR Am Journal Neuroradiol* 2003; 24:658-662

Note that *Invest Radiol* requires only the first three authors followed by "et al." when the authors number is ≥ 4 while the same instruction must be applied for *Radiology, J Magn Reson Imaging (JMRI)* e *AJR Am J Roentgenol (AJR)* when the authors number is ≥ 7 (*Eur Radiol* does not specify). Moreover, *Invest Radiol* requires the journal name in Italic characters followed by a full stop while AJR requires the journal name in Italic characters not followed by a full stop. Finally, *Eur Radiol* and AJR do not want the full stop at the end of each reference.

10.11. «Abstract» and «Keywords»

Your business card to the reader

After having completed the manuscript editing, it is time to write the abstract. It will be, after the title, the business card of your manuscript to the reader, if the paper is published. But do not forget that a reviewer decides whether to review or not review your manuscript after having read the abstract. First, check which type of abstract is required, namely *structured* (e.g. *Radiology* and *AJR Am J Roentegenol*) or *nonstructured* (e.g. *Eur Radiol*). In the second case, a subdivision in four blocks is not explicit but the content is the same. A structured abstract is composed of four sections, not entirely corresponding to the four sections of the main body of the paper we have already discussed (the name of the two first sections can be slightly different in many journals):

Structured or nonstructured Abstract

1. *Purpose*: summarizes the aim of the study (may be similar to the final part of the introduction of the full paper);
2. *Materials and methods*: summarizes the same section of the main body;
3. *Results*: summarizes the same section of the main body;
4. *Conclusion(s)*: summarizes the interpretation of the results (may be similar to the final part of the Discussion of the full paper).

The mandatory limits to the abstract size (e.g. 200-250 words) may appear as a very high hurdle for beginners. Look at the abstracts of papers similar to yours in the same journal and ask for help from more experienced colleagues.

Key words

Some journals require three to five *keywords*, free or to be chosen among predetermined lists, such as the Medical Subject Headings (MeSH), a controlled dictionary for indexing articles on the MEDLINE/PubMed (http://www.nlm.nih.gov/mesh/meshhome.html). Pay attention to the possibility that the keywords you have chosen are used as a criterion for selecting reviewers with a particular cultural background, since one may give a very different evaluation of your manuscript from another.

10.12. Shared Rules

STARD initiative

A detailed presentation of the rules to be respected for a good quality original article on diagnostic performance was done by an important paper [Bossuyt et al, 2003], published in 2003 by *Radiology* and also by: *Annals of Internal Medicine*, *British Medical Journal*, *Clinical Chemistry*, *Journal of Clinical Microbiology*, *The Lancet*, *Nederlands Tijdschrift voor Geneeskunde*. It is a real short manual for checking the quality of a manuscript or an already published paper. An extremely useful checklist is provided to authors so they may avoid omitting important information. The paper is entitled: *Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative*. STARD is an acronym for Standards for Reporting of Diagnostic Accuracy. The authors evaluated 33 papers which proposed a checklist for studies on diagnostic performance. From a list of 75 recommendations, 25 were judged important (Table 10.6). Many of them were discussed in the previous pages of this book. Beginners should edit a manuscript keeping a copy of this list on their table.

Other shared rules are available for articles reporting the results of randomized controlled trials, the CONSORT statement [Moher et al, 2001], recently extended to trials assessing nonpharmacologic treatments [Boutron et al, 2008] or of meta-analyses, the QUOROM statement [Moher et al, 1999]. In particular, systematic reviews and meta-analyses in radiology should evaluate the study validity for specific issues, as pointed out by Dodd et al [Dodd et al, 2004]: detailed imaging methods; level of excellence of both imaging and reference standard; adequacy of technology generation; level of ionizing radiation exposure; viewing conditions (hard versus soft copy).

10.13. Other Recommendations

Read the Instructions for Authors

These can be summarized in the following paragraphs, some of them having been partially discussed above.

Sequence of the items. Apart from the *full title page*, your submission in one or more files should be as follows:

1. Blind title page;
2. Abstract and keywords;
3. Introduction;

Table 10.6. Checklist for the quality control of manuscripts on diagnostic performance according to the Standards for Reporting of Diagnostic Accuracy (STARD)

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS		Describe	
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	The reference standard and its rationale.	
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard.	
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	
	13	Methods for calculating test reproducibility, if done.	
RESULTS		Report	
Participants	14	When study was done, including beginning and ending dates of recruitment.	
	15	Clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered between.	
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Any adverse events from performing the index tests or the reference standard.	
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	
	22	How indeterminate results, missing responses and outliers of the index tests were handled.	
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Estimates of test reproducibility, if done.	
DISCUSSION	25	Discuss the clinical applicability of the study findings.	

From: Bossuyt M, Reitsma JB, Bruns DE et al (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. Radiology 226: 24-28 (with permission of the author and of the copyright owner [RSNA]). This checklist is proposed for use in the practical quality check: in the column on the right side you should put the number of the page of your manuscript where each recommendation has been applied.

4. Materials and methods;
5. Results;
6. Discussion;
7. References;
8. Tables;
9. Captions (or legends) for illustrations.

Each of these items should begin on a new page. All pages must be numbered. Acknowledgments must be submitted in a separate file, in order to keep the origin of the manuscript blinded to reviewers.

Acronyms and abbreviations

Acronyms. Limit their use to those universally recognized. If you used particular acronyms in the abstract (to respect the mandatory number of words), do not use them in the remaining manuscript. Both in abstract and full text, introduce each acronym the first time it appears in the text and then use them systematically. The same rule applies for abbreviations.

Consistency

Consistency. Always use the same word(s) to indicate a concept. Do not worry about repetitions. Always use the same units of measurement (e.g. length always in cm or always in mm). If you list numerical data, always use the same number of decimal places.

Anonimity

Anonymous origin of the manuscript. Systematically avoid making the authors of the manuscript detectable. If you quote one of your own previously published papers, place the word "BLINDED" at the corresponding reference or quote the paper impersonally. Pay attention to avoid making the manuscript origin detectable by the hospital name on the radiologic images in the figures.

Use the adjective "significant" only for statistical significance

"Significant". Never use the adjective *significant* in a non-statistical sense, i.e. to mean the relevance of a result independent from the calculation of a *p* value, or the relevance of a topic in the literature. Strictly limit its use to statistics.

Be cautious

Cautiousness. Avoid self-evaluating your work with excessive and redundant sentences. The affirmation of being the first to have proposed a technique or a procedure should be done using the ritual sentence "*To the best of our knowledge, …*". But you must have done an accurate search on MEDLINE/Pubmed using multiple combinations of keywords with a negative result before claiming to be the first. If you give a negative evaluation of the work done by other authors, do not forget that the object of a criticism is a paper and not its authors (who may be your reviewers…).

10.14. Dealing with the Editor's Response and the Reviewers' Opinions

The Editor's response

After about one or two months (sometimes more than three months) you will receive the response from the editor and the comments of the reviewer(s). The editor's response can be categorized as follows, in decreasing order:

1. acceptance of the manuscript for publication, without any request of modifications;
2. request of slight modifications (*minor revision*);
3. request of important modifications (*major revision*);
4. rejection with offer of *resubmission*;
5. rejection.

The first response is uncommon, but may happen. In cases of the request of minor revision, the probability of final acceptance is very high. This probability is still relatively high in cases of requests of major revision. In both cases, you should pay attention. You must evaluate point by point all the criticisms and suggestions of the editor and reviewers and prepare a new version of the

manuscript (which should be named "R1", where "R" stands for *revision*). Most journals require the submission of:

1. a document answering all the suggestions and criticisms, explaining how you took them into account and the reasons why you decided not to follow a suggestion;

Annotated copy

2. a copy of the manuscript with evidence of the changes introduced, including the erased sentences (use the "revision" function of a word processor). Some journals require the indication of the correspondence between the criticisms of the reviewers and the changes in the text in this "annotated copy";

Clean copy

3. a "clean copy" with the final text, without any evidence of the changes.

One or more requests for revision

We advise trying to answer and prepare an R1 version even after receiving a request for major revisions. However, if you realize that the reviewers have requested changes which cannot be made (e.g. to search for unavailable clinical data) or changes which are overly time-consuming (e.g. to repeat all image evaluations or segmentations, all measurements, or to add more observers), you can opt for submitting the manuscript to another journal, after making the changes according to the suggestions thought to be useful.

Some weeks after the submission of the R1 version, you might receive final acceptance. However, the request of further modifications is relatively common. You will prepare an "R2" version in the same way you did for the R1 version. Then, if you have solved all the problems, you will receive final acceptance.

The offer of resubmission

Sometimes, the rejection of the manuscript is coupled with an offer of *resubmission*. This event is not rare. In a recent analysis of 196 consecutive manuscripts submitted to the AJR Am J Roentgenol, 20 (10%) were accepted, 106 (54%) were rejected, and 70 (36%) were rejected with offer of resubmission [Kliewer et al, 2004].

This offer indicates that the editor, sometimes having a different opinion in comparison with those of the reviewers, thinks that your manuscript is interesting and wants to give you another possibility: resubmitting the paper for a new evaluation cycle, frequently with at least one of the new reviewers remaining the same as the first evaluation cycle. *We advise accepting an offer of resubmission.* In comparison with the request of major revision, you have an advantage and a disadvantage. The advantage is that you do not have either to answer point by point all the criticisms raised by the reviewers or take into account all their suggestions. The disadvantage is having to restart from scratch, with a high probability, if the resubmission is not rejected, of being requested to perform a revision, with a major revision being more likely than a minor one. However, in our experience, resubmission is associated with final success in a good percentage of cases.

Do not lose heart over rejections

The rejection of a manuscript (without any offer of resubmission) is a common event. Do not lose heart. It also happens to experienced scientists submitting a paper of high value. There are many possible reasons for a rejection. Your paper could be flawed by errors in the design, with unrecoverable biases. Another possibility is that the editor and/or the reviewers might not have understood some important aspects of the study, possibly due to a lack of specific knowledge in a particular research field. There is always a random factor in the assignment of the reviewers. In this case, you could write a polite letter

to the editor explaining your disagreement with the reviewers' opinion and asking for the possibility of a resubmission.

Notice that a manuscript can be rejected even in the absence of substantial methodologic criticisms, simply for an evaluation of low priority. In this case, the submission to another journal is perhaps the best option. Remember that you can choose a journal with a higher IF than the previous one and… keep your fingers crossed. Consider also the possibility of submitting the manuscript to a nonradiologic journal. At any rate, trust in yourself. If there are no basic methodologic flaws, your paper will be published in the end.

The final steps

Beginners may think that after one or two rejections or one resubmission, and after R1 and R2 versions, final acceptance has brought an end to the matter. This is not true, even though the additional final steps are usually approached with enthusiasm, knowing that you have almost reached your goal. What are these final steps?

Queries from the editorial office

Firstly, you must answer the queries raised by the editorial office to the authors. Some journals have high quality editorial staff who set the text and understand the meaning in detail. For any problems, they propose modifications by asking the authors for specific approval. Trivial errors are detected in this phase, including a lack of consistency between abstract and text, text and tables, or text and figures, as well as inaccurate references. You may receive as many as 100 queries for a manuscript of regular length. Each of them requires an answer. Some particular problems are discussed between the deputy editor, the staff and you with a series of emails until a good agreed solution is found.

Pay attention to rewording

For authors who are non-native English speakers, the editorial staff may make a rewording/rephrasing of the text to make it more elegant and understandable. Notice that this restyling can be flawed by mistakes in meaning: the new text is in elegant and fluent English but it says something different, sometimes the opposite, from what you want it to mean.

Proofs correction

For these reasons, too, *proofs correction* is a very important task. Carefully check the final text: a shifted colon can entirely change the meaning! To do this job, print the proofs and read all parts of the paper, including authors' names and affiliations, tables (a skipped tabulation may change the entire meaning of the data), and figures (inverted or rotated images, lack of correspondence between images and legends). Do not delegate this job to people who did not directly participate in editing the text. The result could be very bad, and you, as the first or one of the principal authors, will be responsible.

If you realize that possible differences between the proofs and your text cannot be easily detected, the proof correction should be done by two authors. One of them reads the final manuscript, the other – at best, the first author – checks that the proofs are retaining all the original meaning. At any rate, a double reading finds more errors. Remember that in screening mammography the double reading detects about 15% more cancers!

A professional proofs correction should be performed using a system of dedicated symbols which exists in at least three versions: continental European, British, and American [Gustavii, 2003]. For each correction, you should write a symbol on the printed text, repeat it in the margin, writing next to the possible lacking text. At any rate, the corrections should be clearly understandable. The corrected proofs should be faxed to the editorial office. Alternatively, you could use digital proofs correction on the pdf file, sending the corrected pdf file

by email. A further possibility to facilitate (and maybe accelerate) the work of the editorial office is to also submit a list of the corrections in the old form of an "errata corrige", a five-column text containing for each correction: page number; column; line number; text to be changed; and changed text.

Errata-corrige

Let us repeat this once more. Do not underestimate the importance of proofs correction. This is the last possibility of discovering and correcting errors. It can happen that an important error (especially in graphs and tables) is not detected in the first manuscript, nor discovered by the reviewers, and goes unnoticed right up to the final text and the proofs. Only the printed proofs allow the paper to be seen in a new way, hopefully also allowing the error to be detected.

Importance of proofs correction

10.15. To Conclude

What we have described in this book and especially in this last chapter explains why performing a study and writing a paper requires much more time than what is thought by those not involved in this type of work. From the ideation of a prospective study on diagnostic performance to its publication the time is measured in years.

Long times

An example calculation:

1. conception and initial discussion among colleagues = 2 months;
2. writing and editing the proposal to be submitted to the Institutional Review Board = 1-2 months;
3. approval by the Institutional Review Board = 2 months;
4. patient enrollment = 6-12 months;
5. data acquisition and analysis = 2-3 months;
6. manuscript editing and online submission = 3 months;
7. waiting time for the journal response = 2-3 months;
8. editing of answers to reviewers and editing of the R1 version= 1-2 months;
9. return of proofs to the authors and proofs correction = 3 months;
10. waiting time for final publication = 2-4 months;

Overall time = 24-36 months = 2-3 years.

Sometimes online first publication may shorten this time by several months, but experienced people agree that the preceding time evaluations are largely optimistic. In cases of randomized controlled trials, almost all phases are indubitably longer.

Is it worth it?

So what now? Is it all worth it? When we compare the working time with the results, this question arises spontaneously. The answer is subjective. We say yes, for reason and passion.

For reason, because physicians (for instance, radiologists) who are also researchers have a higher degree of clinical knowledge and are able to offer a superior degree of diagnosis and treatment to patients, which is our final mission.

For passion, a passion for a world game where we can interact with the best specialists in each research field, the reviewers of the top journals. We can obtain their evaluation of our work and exchange opinions with them, making real science in the real world.

In ending his autobiography, Luca Cavalli-Sforza stated:

In my opinion, my need to be always active is like being a child who is relentlessly playing, sometimes changing the game. Of course, the reader thinks that I can do it because making science is basically equivalent to playing. It really is a game, in the sense that it engages the researcher as a game does. However, it is different due to its long-term special purpose [Cavalli-Sforza, 2005].

Making science is a remarkable long-term game. We love it both for the intellectual knowledge and pleasure it brings, which go beyond any ambition or possibility available in an academic career, and for its effect of improving the quality of clinical medicine.

A final hope

Every game has its rules. We hope that this book contributes to explaining some of the rules of scientific research in radiology.

References

Bossuyt M, Reitsma JB, Bruns DE et al (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. Radiology 226:24-28

Boutron I, Moher D, Altman DG, et al for the CONSORT Group (2008) Methods and processes of the CONSORT Group: example of an extension for trials assessing nonpharmacologic treatments. Ann Intern Med 148:W60-66

Cavalli-Sforza L, Cavalli-Sforza F (2005) Perché la scienza. L'avventura di un ricercatore. Milan. Mondadori

Dodd JD, MacEneaney PM, Malone DE (2004) Evidence-based radiology: how to quickly assess the validity and strength of publications in the diagnostic radiology literature. Eur Radiol 14:915-922

Gustavii B (2003) How to write and illustrate a scientific paper. New York. Cambridge University Press

Kerkut GA (1983) Choosing a title for a paper. Comp Biochem Physiol 74A:1. Quoted in: Gustavii B (2003) How to write and illustrate a scientific paper. New York. Cambridge University Press

Kliewer MA, DeLong DM, Freed K et al (2004) Peer review at the American Journal of Roentgenology: how reviewer and manuscript characteristics affected editorial decisions on 196 major papers. AJR Am J Roentgenol 183:1545-1550

Moher D, Cook DJ, Eastwood S et al (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 354:1896-1900

Moher D, Schulz KF, Altman DG (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357:1191-1194

Rogers SM (2007) Mastering scientific and medical writing. Berlin. Springer

Sardanelli F, Podo F, D'Agnolo G et al (2007) Multicenter comparative multimodality surveillance of women at genetic-familial high risk for breast cancer (HIBCRIT study): interim results. Radiology 242:698-715

World Medical Association (2004) Helsinki Declaration. http://www.wma.net/e/policy/b3.htm. Accessed February 19, 2008

Subject and Noun Index

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y