Stephen P. Glasser *Editor*

Essentials of Clinical Research

Second Edition

Springer

Essentials of Clinical Research

Stephen P. Glasser
Editor

Essentials of Clinical Research

Second Edition

Springer

Editor
Stephen P. Glasser
Division of Preventive Medicine
University of Alabama at Birmingham
Birmingham, AL, USA

ISBN 978-3-319-05469-8 ISBN 978-3-319-05470-4 (eBook)
DOI 10.1007/978-3-319-05470-4
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014940938

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
The Beginning – Historical Aspects of Clinical Research, Clinical Research: Definitions, "Anatomy and Physiology," and the Quest for "Universal Truth"

Stephen P. Glasser

> *Scientific inquiry is seeing what everyone else is seeing,*
> *but thinking of what no one else has thought*
>
> A. Szentgyorgyi. 1873 (he won the Nobel Prize for isolating Vitamin C) [1].

Abstract To answer many of their clinical questions, health care practitioners need access to reports of original research. This requires the reader to critically appraise the design, conduct, and analysis of each study and subsequently interpret the results. This first chapter reviews some of the key historical developments that have led to the current paradigms used in clinical research, such as the concept of randomization, blinding (masking) and, placebo-controls.

Keywords Clinical research definition • Clinical research history

Introduction

As a former director of a National Institutes of Health (NIH)-funded K30 program it was my responsibility to provide a foundation for young researchers to become independent principal investigators. A part of our curriculum was a Course entitled

Hulley S, Cummings S, Browner WS. Designing clinical research. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 2000.

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_1,

'The Fundamentals of Clinical Research.' This course, in addition to guiding students, towards becoming research investigators, was also designed to aid 'students' who wanted to read the medical literature more critically. The importance of this latter point is exemplified by the study of Windish et al., who note "*physicians must keep current with the clinical information to practice evidence-based medicine.... to answer many of their clinical questions, physicians need access to reports of original research. This requires the reader to critically appraise the design, conduct, and analysis of each study and subsequently interpret the results*" [2]. Although aimed at physicians, this observation can and should be applied to all health scientists who must read the literature in order to place the results in context. The Windish study surveyed 277 completed questionnaires that assessed knowledge about biostatistics, and study design. The overall mean percent correct on statistical knowledge and interpretation of results was 41.4 %.

It is my belief that the textbooks currently available are epidemiologically "slanted". There is nothing inherently wrong with that slant, but I have written this book to be more specifically geared to the clinical researcher interested in conducting Patient Oriented Research (POR). In this first chapter I will provide a brief overview of the history of clinical research. The chapter will also address the question of why we do clinical research; define 'clinical research'; discuss our quest for 'universal truth' as the reason for doing clinical research; outline the approach taken to answer clinical questions; and describe (as Hulley and colleagues so aptly put it) 'the anatomy and physiology of clinical research' [3].

Future chapters will examine such issues as causality (i.e., causal inference or cause and effect relationships); the strengths and weaknesses of the most popular clinical research designs; regression to the mean; clinical decision making; meta-analysis; and the role of the Food and Drug Administration (FDA) in the clinical trial process. We will also focus on issues related to randomized clinical trials, such as the intention-to-treat analysis, the use and ethics of placebo-controlled trials, and surrogate and composite endpoints.

Definition of Clinical Research

The definition of clinical research might appear to be self-evident; however, some researchers have narrowly defined clinical research to refer to clinical trials (i.e., intervention studies in human patients), while others have broadly defined it as any research design that studies humans (patients or subjects) or any materials taken from humans. This latter definition may even include animal studies, the results of which more or less directly apply to humans. For example, in 1991, Ahrens included the following in the definition of clinical research: studies on the mechanisms of human disease; studies on the management of disease; in vitro studies on materials of human origin; animal models of human health and disease; the development of new technologies; the assessment of health care delivery; and field surveys [4]. In an attempt to simplify the definition, some wits have opined that clinical research occurs when the individual performing the research is required to have malpractice insurance, or when the investigator and the human subject are, at some point in the study, in the same room, and both are alive and warm. So, there is a wide range of definitions of clinical research, some valid, some not. I have chosen to adopt a 'middle of the road' definition that encompasses the term 'patient-oriented-research,' which is defined as research conducted with human subjects (or on material of human origin) for which the investigator directly interacts with the human subjects at some point during the study. It is worth noting that this definition excludes in vitro studies that use human tissue that may or may not be linked to a living individual unless the investigator during the conduct of the trial has significant interaction with a living breathing human.

History of Clinical Research

Perhaps the first clinical trial results were those of Galen (circa 250 BC) who concluded that 'some patients that have taken this herbivore have recovered, while some have died; thus, it is obvious that this herbivore fails only in incurable diseases.' Galen's observations underline the fact that even if we have carefully and appropriately gathered data, there are still subjective components to its interpretation, indicating our quest for 'universal truth' may be bedeviled more by the interpretation of data than by its accumulation (more about this in Chap. 3).

James Lind is generally given credit for performing and reporting the first 'placebo-controlled' interventional trial in the treatment and prevention of scurvy. In the 1700s, scurvy was a particularly vexing problem on the long voyages across the Atlantic Ocean. The research question that presented itself to Lind was how to prevent the condition. To arrive at an answer, Lind did what every good researcher should do as the first step in converting a research question into a testable hypothesis—he reviewed the existent literature of the time. In so doing, he found a report from 1600 that stated '*1 of 4 ships that sailed on February 13th, 1600, was fortuitously supplied with lemon juice, and almost all of the sailors aboard the one ship were free of scurvy, while most of the sailors of the other ships developed the disease*.' This was not a planned experiment, however. The first planned experiment was perhaps one that involved smallpox, performed in 1721, in which six inmates of Newgate Prison were offered to have their sentence commuted if they volunteered for inoculation. All remained free of smallpox. However, in this experiment there was no concurrent control group. Returning to Lind's review of the literature, on the one hand, Lind's job was easy; there was not a great deal of prior published works. On the other hand, Lind did not have computerized searches via Med Line, Pub Med etc available.

As a result of the above, in 1747, Lind set up the following trial. He took 12 patients 'in the scurvy' on board the HMS *Salisbury*. '*These cases were as similar as I could have them…. They lay together in one place…and had one diet common to all. The consequence was that the most sudden and visible good effects were perceived from the use of oranges and lemons*.' Indeed, Lind evaluated six treatment groups:

Table 1.1 Lind's 1747 "clinical trial"

Lind's description	Modern day RCT correlate
"These cases were as similar as I could find them"	Inclusion/exclusion criteria
"They lay together in one place and had one diet common to all"	Common treatment save for the intervention of interest
"Six treatment groups were evaluated"	Parallel group design
"The rest served as controls"	Active control groups
"Two…were put under a course of sea water"	Placebo group?
"The… the most sudden and visible good effects were perceived from oranges and lemons"	Interpretation

'*one group of two was given oranges and lemons. One of the two recovered quickly and was fit for duty after 6 days, while the second was the best recovered and was assigned the role of nurse for the remaining patients*.' The other groups were each treated differently and served as controls. If we examine Lind's 'study' we find a number of insights important to the conduct of clinical trials as follows. For example, he noted that *'on the 20th May, 1747, I took twelve patients in the scurvy on board the Salisbury at sea… Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees…' Here Lind was describing eligibility criteria for his study. He continues, '…They lay together in one place, being a proper apartment for the sick in the forehold; and had one diet in common to all…' '… Two of these were ordered each a quart of cyder a day. Two others took twenty five gutts of elixir vitriol three times a day upon an empty stomach,*

… Two others took two spoonfuls of vinegar three times a day
… Two … were put under a course of sea water.
… Two others had each two oranges and one lemon given them every day.
… The two remaining patients took the bigness of a nutmeg three times a day.'

By this description, Lind described the interventions and controls. To continue, '… *The consequence was that the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them being at the end of six days fit four duty. The spots were not indeed at that time quite off his body, nor his gums sound; but without any other medicine than a gargarism or elixir of vitriol he became quite healthy before we came into Plymouth, which was on the 16th June*.' This latter description represents the outcome parameters and interpretation of his study. In summary, Lind addressed the issues of parallel-group design and the use of control groups, and he attempted to assure similarity between the groups except for the intervention (Table 1.1).

Clearly, sample size considerations and randomization were not used in Lind's trial nor were ethics and informed consent mentioned, but this small study was amazingly insightful for its time. Other selected milestones in the history of clinical research include:

- Fisher's introduction of the concept of randomization in 1926; [5]
- The announcement in 1931 by the Medical Research Council that they had appointed '*a therapeutics trials committee…to advise and assist them in arranging*

for properly controlled clinical tests of new products that seem likely on experimental grounds to have value in the treatment of disease' [6];

- Amberson and colleagues' introduction of the concept of 'blindness' in clinical trials [6], and their study of tuberculosis patients where the process of randomization was applied [7]. They noted that after careful matching of 24 patients with pulmonary tuberculosis, the flip of a coin determined which group received the study drug [7].

Further analysis of the tuberculosis streptomycin study of 1948 is regarded by many, as the beginning of the beginning of the modern era of clinical research and is instructive in this regard. In the 1940s tuberculosis was a major public health concern, and randomization was being recognized as a pivotal component to reduce bias in clinical trials [8]. As a result the Medical Research Council launched a clinical trial in which 55 patients were randomized to treatment with bed rest (the standard of care treatment at that time) and 52 were treated with bed rest alone [9].

Other significant developments include reference to the use of saline solution in control subjects as a placebo, and the requirement in 1933 that animal toxicity studies be performed before human use [8]. In the 1940s, the Nuremberg Code, the Declaration of Helsinki, the Belmont Report, and the doctrine of Good Clinical Practice (GCP) were developed, which will be discussed in more detail later. As mentioned above, In1948, the Medical Research Council undertook a streptomycin study [9] which was perhaps the first large-scale clinical trial using a properly designed randomized schema. This was followed by an antihistamine trial that used a placebo arm and double-blind (masked) design [10].

In 1954, there were large-scale polio studies—field trials of 1.8 million school-age children. A controversy regarding the best design resulted in two trials, one design in which some school districts' second graders received the dead virus vaccine while first and third graders acted as the controls (i.e. a group clinical trial); and another design in which second graders randomly received either the vaccine or a saline injection. Both studies showed a favorable outcome for the vaccine (Fig. 1.1).

In 1962, the thalidomide tragedy became widely known and resulted in the tightening of government regulations as they applied to drug development and approval (also see Chap. 6). The story behind this tragedy is instructive. By 1960, thalidomide worldwide was being sold, but not in the United States. At the time, the prevailing US law was the 1938 Federal Food, Drug, and Cosmetic Act, which required proof of safety be sent to the FDA before a medication could be approved for sale in the United States. The law did not require demonstration of efficacy for approval. It also allowed "investigational" or "experimental" use of a drug while approval for its sale was being sought, allowing a medication to be widely distributed prior to approval. The application for use of thalidomide in the USA was given to Frances Kelsey who noted a lack of teratogenicity data, and she also had other worries about thalidomide. As a result, Kelsey rejected the application and requested additional data from the company, who complained to her superiors that she was nit-picking and unreasonable. Kelsey continued to refuse to allow thalidomide for sale in the United States, and in total, the company resubmitted its application to

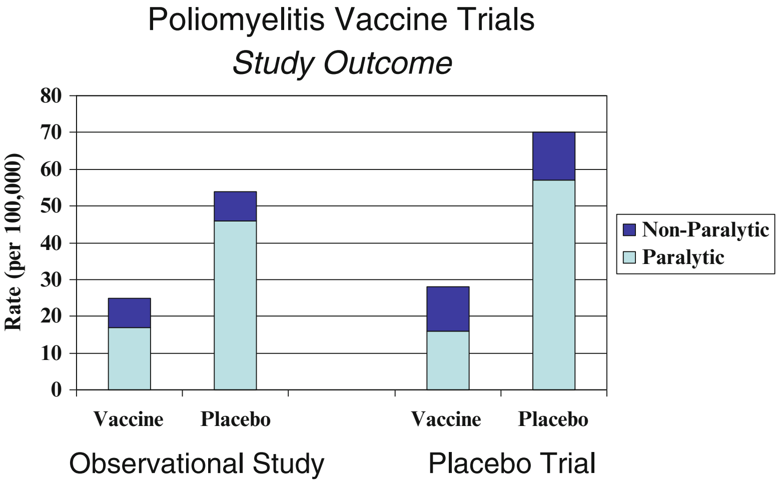


Fig. 1.1 Results from the use of polio vaccines used in both an observational trial and a placebo controlled clinical trial

the FDA six times, but with no new evidence in those applications, Kelsey refused approval. Subsequently, reports regarding a number of birth defects were reported and the drug was subsequently removed worldwide [11].

As prior mentioned, at the time of the thalidomide disaster, trials of new drugs were required to prove safety but not efficacy as described under the FDA's 1938 Act. As a result of the disaster, tightening of the regulations was instituted and trials were to have an "*adequate and well-controlled design*" before approval of any new drug. This was followed by the Drug Efficacy Study Implementation (DESI) review and the FDA's development of the four stages of clinical trials necessary for new drug approval, which set the stage for today's drug approval process (see Chap. 6).

In the 1970s and 1980s, clinical research was prospering, but by the 1990s there began a decline in the number of new clinical investigators. This trend caught the eye of a number of academicians and the NIH, which then commissioned the Institute of Medicine (IOM) to address ways to stimulate individuals to pursue careers in clinical investigation, to define appropriate curricula for training, and to ensure adequate support mechanisms for retaining clinical researchers.

The NIH also developed granting mechanisms for supporting individual clinical investigators at various levels of their careers (e.g. K23 and K24 grants) and for programmatic support of institutions that developed clinical research training programs (K30 grants), and most recently its establishment of Centers for Clinical and Translational Science (CCTS). The IOM report documented the decline in clinical investigators (particularly MD investigators), and noted that the time commitment necessary to do clinical research was underappreciated [12].

DeMets and Califf more recently noted, '*we are entering an era in which the imperative to understand the rational basis for diagnostic and therapeutic options has become a major force in medical care.*' Medical products (drugs, devices, and biologics) are proliferating simultaneously with substantial restructuring of the delivery of health care, with a focus on evidence to support medical intervention [13].

Today, we are left with the 'good, the bad, and the ugly' regarding clinical research. The 'good' is that many experts think that sound comprehension of the scientific method and exposure to biomedical research comprise the essential core of medical education, and that the very essence of the American academic model is a balance between education, patient care, and research. The 'bad' is the increasing number of voices questioning the relevancy of research in academic health centers, as well as those concerned about the commitment to other components of training and the cost of research in a setting where the 'triple threat' (i.e., excelling in teaching, patient care, and research) may no longer be tenable given the increasing complexity of each area. The 'ugly' is that in 2003 only about 3 cents of every health care dollar was spent on medical research (more recently this has dropped to 2 cents); and, it was estimated that only 5 % of Congress could be counted on to take the initiative and be leaders in the support of clinical research; and few potential investigators were being supported to pursue careers or were given enough time to conduct research. By and large, these same issues persist today. In addition, today's challenges add even greater burdens to clinical research. It is generally believed that today's studies cost too much, fail to recruit adequate numbers of subjects/patients into trials, fail to start in a timely fashion, may not even be asking the correct questions or studying the correct endpoints, and study results are often not published (publication bias, is an issue here). In fact, Pfizer has reported that recently, 60 % of the total drug development costs go to conducting clinical trials, compared to 30 % in the 1980s. These increased costs (which are 1.5–3x higher than many other countries) are making the US less competitive worldwide.

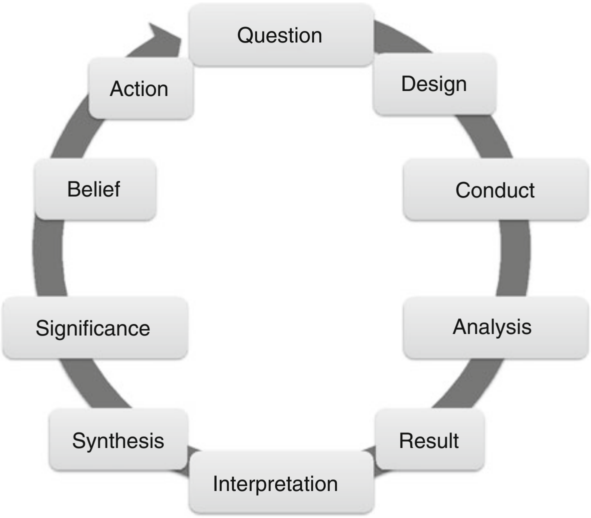
Our Quest for Knowledge

With the above background, how do we begin our quest for knowledge? In general, research questions are generated in a variety of settings (e.g., during journal reading, hospital rounds, discussions with colleagues, seminars, and lectures). The resultant questions can then be refined into a research idea and, after further review of the literature, ultimately developed into a hypothesis. Based on a number of factors (to be discussed in subsequent chapters), a study design is chosen, and the study is then preformed and analyzed, the results of which are then interpreted and synthesized. These results add to the body of knowledge, and this may raise additional questions that will invariably generate further research (Fig. 1.2).

Of course, the primary goal of clinical research is to minimize presumption and to seek universal truth. In fact, in science, little if anything is obvious, and the interpretation of results does not mean truth, but is really an opinion about what the results mean. Nonetheless, in our quest for universal truth, Hully and colleagues have diagrammed the steps that are generally taken to seek this 'truth' (Fig. 1.3) [3].

Much of research is to explore this concept of opening ones mind. That is, *"to know that we know what we know, and that we do not know what we do not know, that is*

Fig. 1.2 Scientific method paradigm



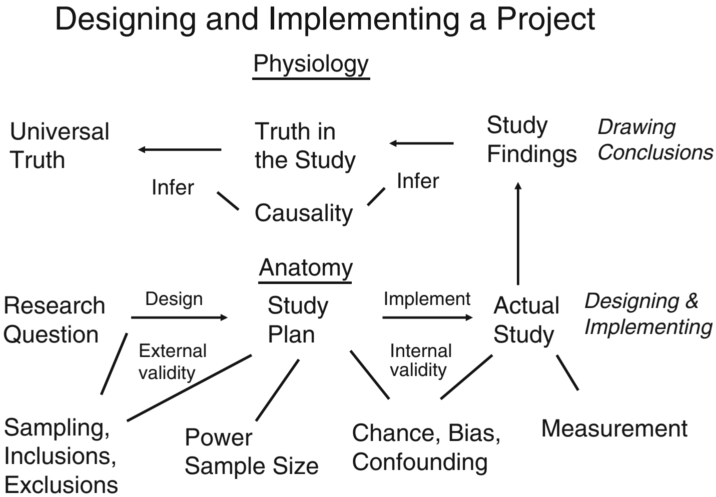


Fig. 1.3 A schematic of the design and implementation of a study

true knowledge (Henry David Thoreau (1817–1862))", or "scientific inquiry is seeing what everyone else is seeing, but thinking of what no one else has thought" (A. Szentgyorgyi. 1873 won the Nobel Prize for isolating Vitamin C). Most (perhaps all) people generally know what they know and know what they do not know. What get's most of us in trouble is that we do not know what we do not know (Fig. 1.4), and the largest "piece of the pie" falls in the last category.

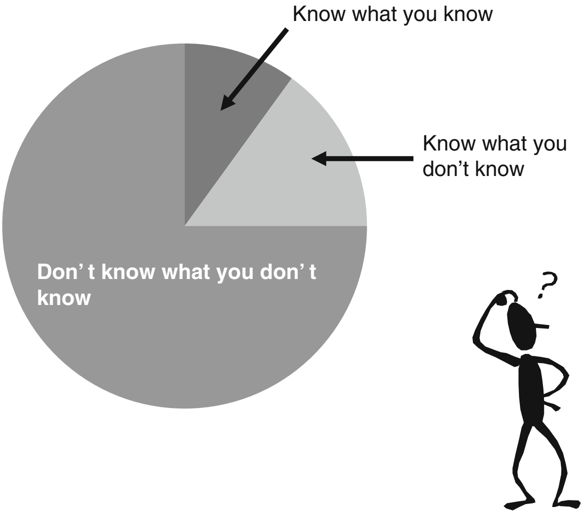


Fig. 1.4 One's universe of knowledge

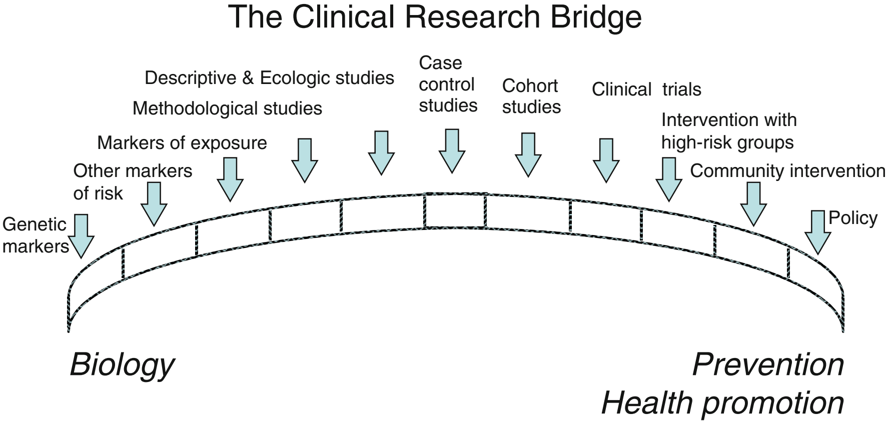


Fig. 1.5 Portrays the broad range that encompasses the term "clinical research"

This is exemplified, by considering the question of what the "experts" in the past really knew. Consider the following quotes:

"A journey such as that envisioned by Columbus is impossible. Among the many reasons that can be cited as to the folly of this enterprise is the well known fact that the Atlantic Ocean is infinite and therefore impossible to traverse"
(From a committee report to King Ferdinand and Queen Isabella, 1486)
"Who the hell wants to hear actors talk?"
From Jack Warner, Warner Bros. Pictures, 1927
"I think there is a world market for about 5 computers"
From: TJ Watson, CEO of IBM, 1943

"There is no reason for any individual to have a computer in their home."
From: Ken Olsen, President of Digital Corporation, 1977

Finally, it should be realized that clinical research can encompass a broad range of investigation as portrayed in Fig. 1.5.

References

1. http://www.brainyquote.com/quotes/authors/a/albertszentgyorgyi.html
2. Windish DM, Hoot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA. 2007;298:1010–22.
3. Hulley S, Cummings S, Browner WS. Designing clinical research. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 2000.
4. Ahrens E. The crisis in clinical research: overcoming institutional obstacles. New York: Oxford University Press; 1992.
5. Fisher R. The design of experiments. Edinburgh: Oliver and Boyd; 1935.
6. Hart PD. Randomised controlled clinical trials. BMJ. 1992;302:1271–2.
7. Amberson JB, MacMahon BT, Pinner M. A clinical trial of sanocrysin in pulmonary tuberculosis. Am Rev Tuber. 1931;24:401–35.
8. Hill AB. The clinical trial. Br Med Bull. 1951;7:278–82.
9. White L, Tursky B, Schwartz G. Placebo: theory, research, and mechanisms. New York: Guilford Press; 1985.
10. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ. 1948;ii:769–82.
11. Thalidomide. http://en.wikipedia.org/wiki/Thalidomide
12. Institute of Medicine. Careers in clinical research: obstacles and opportunities. Washington, DC: National Academy Press; 1994.
13. DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials: Part I. Circulation. 2002;106:746–51.

Chapter 2
Introduction to Clinical Research Concepts, Essential Characteristics of Clinical Research, Overview of Clinical Research Study Designs

Stephen P. Glasser

> *To educate is to guide students on an inner journey toward more truthful ways of seeing and being in the world.*
>
> (Parker J. Palmer) [1]

Abstract This chapter addresses some of the central concepts related to clinical research such as sampling, hypothesis generation, and what is meant by the strength of scientific evidence. We also begin to discuss the different clinical research designs along with their respective strengths and weaknesses.

Keywords Sampling • Hypothesis • Prospective and retrospective cohort design • Case-control design • Case cohort design • Cross-sectional design • Type I and type II error

Principles for the conduct of research are set forth in internationally recognized documents such as the Declaration of Helsinki and the Guideline for Good Clinical Practice (GCP) of the International Conference on Harmonization (ICH-see Chap. 6). The principles of these and other standards are translated into legal requirements through laws and regulations that are enforced by national authorities such as the US FDA (see Chap. 6). The issues addressed by GCP include such things as protecting research subjects, ensuring objectivity in research, communication information about clinical trials, informed consent, and the very conduct of clinical trials including independent review and safety monitoring. In recent years clinical research has been discussed in the lay media, and this has (mostly) negatively impacted recruitment (also see Chap. 8).

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_2,

Sampling

An essential characteristic and the goal of any clinical research is to make inferences from the population under study (the sample or study population) and apply those inferences to a broader population (the target population i.e. the population about which we want to draw conclusions). Imagine if the investigator could only learn about and apply the results in the sample population? Rather we must be able to extrapolate the results of the findings in the sample population to a broader group of patients-otherwise the results would have little utility (Fig. 2.1). Thus, one of the most important weaknesses of any study is that inferences drawn from a study are based on a limited sample (again, a sample is a select subset of a population that the investigator hopes represents the general population perfectly, but which is unlikely to do so). This aforementioned limitation is further compounded by the fact that disease is not distributed randomly, whereas samples tend to be, and that the causes of disease are multifactorial. Thus, ideally, when performing clinical research, we would like to include everyone in our study who has the disease of interest. Because this is impossible we settle for a sample of the diseased population, however, the researcher now has to deal with a degree of uncertainty (see Chap. 18). Because different samples contain different

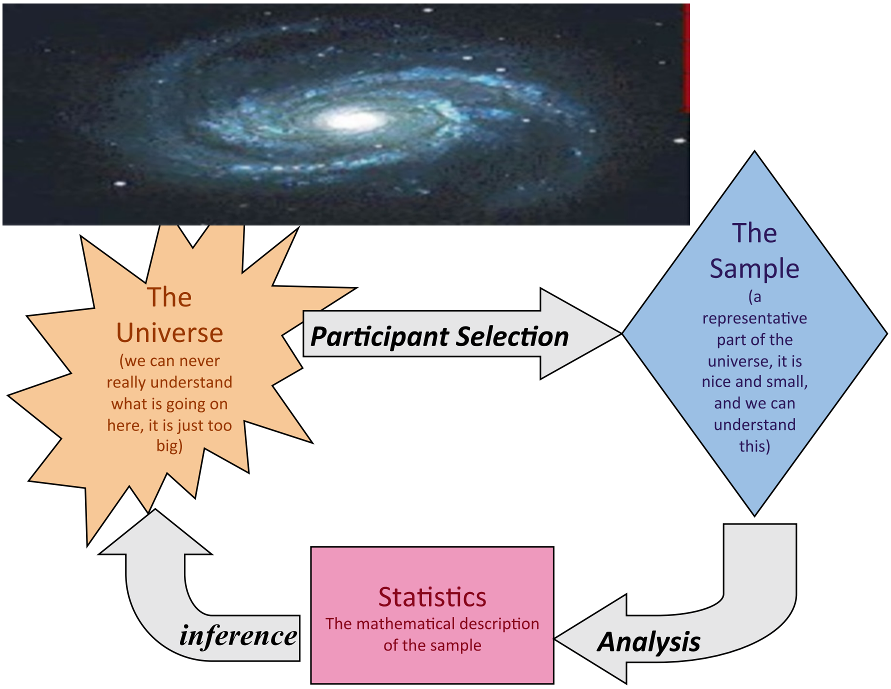

The Universe
(we can never really understand what is going on here, it is just too big)
Participant Selection
The Sample
(a representative part of the universe, it is nice and small, and we can understand this)
Analysis
Statistics
The mathematical description of the sample
inference

Fig. 2.1 The sample and how it relates to the universe

Table 2.1 Potential sampling errors

Selecting study participants
Selection bias
Non-respondent bias:
Volunteer or referral bias
External validity
Sampling bias
Ascertainment bias
Prevalence-incidence bias
Berkson bias
Healthy worker effect
Detection bias: the risk factor investigated itself may lead to increased
Diagnostic
Overmatching bias

people with different co-morbidities, and differing experiences, we end up with different data. The question now facing the researcher is which data from which sample is most representative of the entire population? Sampling errors commonly result in type I and type II errors. For example, if the researcher finds a certain effect of an interventional therapy, the question to be asked is 'how likely is it that this therapy observation that was made from this sample is falsely representing the total population (that is the intervention in the sample population shows no effect, but if the total population had been exposed to the intervention there would have been an effect)? This potential false result is a type II error. The reverse situation is a total population would in fact have a therapy effect, but the sample studied shows no such effect. This is a type I error and is reflected by the p value.

Sampling bias is also a major problem (Table 2.1). For example, considering who responds to certain types of advertisement to recruit subjects can bias the sample. If random digit telephone dialing is used, subjects who do not have a phone cannot be recruited, if newspaper advertisement is utilized people who do not read newspapers cannot respond, etc.

A suggested solution to the sampling issue is to use random sampling; but, random sampling does not guarantee 'good' sampling. As an example, consider If you draw repeated random samples of size 100 and 1,000 from a population with 50 % women the largest and smallest number of women in a sample of 100 can range from 33 to 68 and in a sample of 1,000 from 450–550.

The Linear-Semilinear Relationship of Biological Variables

Another important concept of clinical research is the fact that most, if not all biological variables have a linear–semilinear relationship in terms of exposure and outcomes, whereas clinical medicine is replete with the use of 'cut-points' to separate

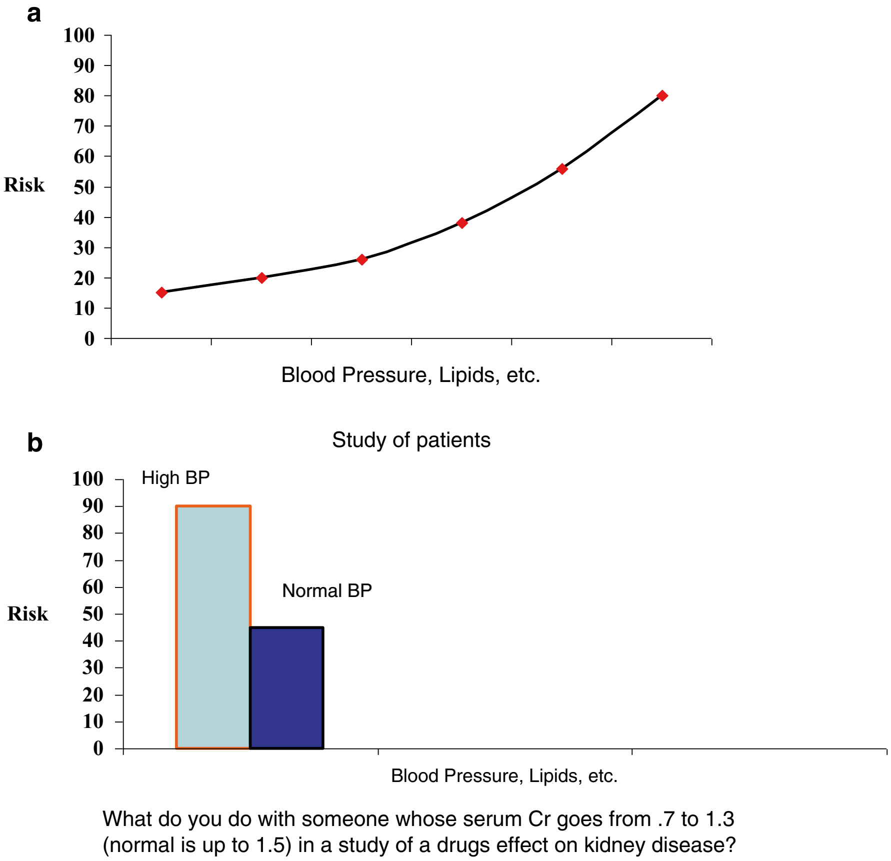


Fig. 2.2 A comparison of the epidemiologic way of determining risk (**a**) vs. the clinical way (**b**)

normal and abnormal or effect and no effect (Fig. 2.2a, b). A cut-point presumes that there is some value or range of values that separates normal form abnormal rather than considering that the relationships tend to be linear.

Strength of Relationships

Additionally, clinical research relates to what we mean when we talk about 'the strength of evidence.' The greatest strength of evidence is often attributed to the randomized clinical trial (RCT). In fact, in response to the question of what is the best clinical research design, the answer generally given is 'the RCT,' when in fact the correct answer should be 'it depends,' an answer which will be further discussed later in this book. What is actually meant by 'the highest level of evidence' is how certain we are that an exposure and outcome are causally related, that is, how certain we are that an effect is the result of a given cause, and that the observations do not just reflect that an association exists; but, that they are not causally related.

The Hypothesis

Let's return to the question: 'What is the best study design?' This is a different question from 'What is the best study design for a given question, and given the specific question, which study design leads to the highest level of evidence?'; which may finally be different from asking 'What is the study design for a given question that will result in the greatest certainty that the results reflect cause and effect?' This latter question is really the one that is most often sought, and is the most difficult to come by (see Chap. 16). Other important factors in considering the most appropriate study design, besides the most important factor—ethics—include the natural history of the disease being studied, the prevalence of the exposure, disease frequency, the characteristics and availability of the study population, measurement issues, and cost.

Let us now return to our quest for 'universal truth.' What are the steps we need to take in order to achieve 'truth'? The fact is that truth is at best elusive and is not actually achievable since truth is more a function of our interpretation of data, which is in part dictated by our past experiences, than any finite observation that is absolute. The steps needed to achieve this uncertain quest for truth begins with a research question, perhaps the result of a question asked during teaching rounds, or stimulated by contact with a patient, or provoked during the reading of a book or journal, and so on. The research question is usually some general statement such as 'Is there an association between coffee drinking and myocardial infarction (MI)?' or 'Is passive smoke harmful to a fetus?' Let us examine this last research question and consider its limitations in terms of a testable hypothesis. In addressing a question such as 'Is passive smoke harmful to a fetus?' one needs first to ask a few questions such as: 'what is the definition of 'harmful'; how will passive smoke be measured and what do we mean by the term i.e. how is it to be defined in the study to be proposed?' Answering these questions comes nearer to something that is testable and begins to define the clinical research design that would have the greatest level of evidence with that specific question in mind. For the question proposed above, for example, it would be best, from a research design perspective, to randomize exposure of pregnant women to both passive smoke and 'placebo passive smoke.' But considering the ethics issue alone, this would not be acceptable; thus, an RCT would not be the 'best study design' for this research question, even if it would lead to the 'highest level of evidence'.

The hypothesis is generally (for the traditional approach of superiority testing) stated in the null (Ho). The alternative hypothesis (H_A) i.e. the one you are really interested in is, for example, that a new drug is better than placebo. That is, if one wants to compare a new investigational drug to placebo, the hypothesis would be constructed in the null, i.e. that there is no difference between the two interventions. If one rejects the null, one can then say that the new drug is either better (or worse-depending on the results of the study) than placebo. By the way, if the null is not rejected one cannot say that the new drug is the same as placebo, one can only claim that no difference between the two is evident from these data (this is more than a nuance as will be discussed later).

In order to understand why the hypothesis is stated in the null and why one cannot accept the null but only reject it, consider the following three examples (taking a trip with your family, shooting baskets with Michael Jordon, and contemplating the US legal system). Consider the scenario outlined by Vickers [2] where you have just finished packing up your SUV (a hybrid SUV no doubt) with all of your luggage, the two kids, and your dog, and just as you are ready to depart; your wife says ‘honey, did you pack the camera?’ At least two answers present themselves; one that the camera is in the automobile, or two that the camera is in the house. Given the prospect of unpacking the entire SUV, you decide to approach the question with, ‘the camera is not in the house (Ho) i.e. it is in the car’. If you in fact do not find the camera in the house you have rejected your null and your assumption is that it is in the car. Of course, one can easily see that the camera could be in the house (you just did not find it), and even if you did such a thorough job of searching the house that you can be almost certain that it is not there, it still may not be in the car (you might have left it elsewhere (the office, a prior vacation, etc.)) Another way to look at this issue is to envision that you are out on the basketball court when Michael Jordon comes in. You challenge him to a free throw shooting contest and he makes 7 of 7 while you make 3 of 7. It turns out the p value for this difference is 0.07 i.e. there is no “statistically significant difference between the shooting skills of MJ and your shooting skills” you can draw your own conclusions about this likelihood [2]. In the Woman’s Health Initiative (WHI), women eating a low fat diet had a 10 % reduction in breast cancer compared to controls P = .07. This was widely interpreted, as low fat diets don’t work. In fact, the NY Times trumpeted that ‘low fat diets flub a test’ and that the study provided ‘strong evidence that the war against all fats was mostly in vain’. This is what we call accepting the null hypothesis (i.e. it was not rejected so it was accepted) and is to be avoided i.e. failure to reject it does not mean you accept it, rather it means that these data do not provide enough evidence to reject it. By the way, guess what happens when the next study does reject the null-‘but they said it did not work!’.

Finally, consider our Anglo-American legal system. *It is no mere coincidence that the logic of hypotheses testing in scientific inquiry is identical to that which evolved in the Anglo-American legal system and most of the following descriptions are taken from The Null Logic of Hypothesis Testing found on the World Wide Web* [3]. *Much of the pioneering work in the logic of hypothesis testing and inferential statistics was done by English mathematicians and refined by their American counterparts. For instance consider the contributions made by W.S. Gossett, R.A. Fisher, and Karl Pearson to the logic of hypothesis testing and statistical inference. The concept of the null hypothesis can be compared to the legal concept of guilty* *vs. non guilty, the latter of which does not mean innocence. What is interesting is that the guilt vs. innocent scenario involves two diametrically opposed logics, one affirmative and the other null. From the time a crime is reported to the police an affirmative, accusatory, and inductive logic is followed. Detective X gathers the evidence, follows the evidentiary trail, and based upon the standard of probable cause, hypothesizes that the accused is guilty and charges him accordingly. The District Attorney reviews the case for probable cause and quality of evidence and affirms the accusation. The case is argued affirmatively before the grand jury, and they concur. But relative to the jury, at the point the trial begins, the logic is reversed, it is no longer affirmative, it becomes null. The jury, the trier of the facts, is required to assume that the defendant is not guilty unless the facts established otherwise. Let's abstract this two part logical process and represent it symbolically. The police, the prosecutor, and the grand jury hypothesized (H_A) that the accused (X) committed the crime (Y).*

$$H_A : (X \rightarrow Y)$$

The jury on the other hand hypothesizes (H_0) that the accused (X) was not guilty of the crime (Y) unless the evidence reached the standard of "beyond a reasonable doubt".

$$\boxed{H_0: (X \not\rightarrow \cdots\cdots Y)}$$

Formulating the logic in this manner, one can be certain of three things. Either:

H_0 is true, the accused is not guilty, or
H_A is true, accused is guilty,
and
H_0 and H_A cannot both be true.

The logic of establishing someone's guilt is not the simple converse of the logic of establishing his/her innocence. For instance, accusing someone of a crime and requiring them to prove their innocence requires proving a negative, something that is not logically tenable. However, assuming that someone is not guilty and then assessing the evidence to the contrary is logically tenable (Fig. 2.3).

The decision matrix in Table 2.1 shows the possible outcomes and consequences of this legal logic as applied to the case of the accused, our hypothetical defendant. Assume H_0: the accused is not guilty unless the evidence is convincing beyond a reasonable doubt. Notice that in terms of verdicts and outcomes, there are two kinds of errors the jury might have made, identified as (I) and (II).

Type I Error The jury finds the accused guilty when in fact he is not guilty.
Type II Error The jury finds the accused not guilty when in fact he is guilty.

Compare this with the Table 18.2

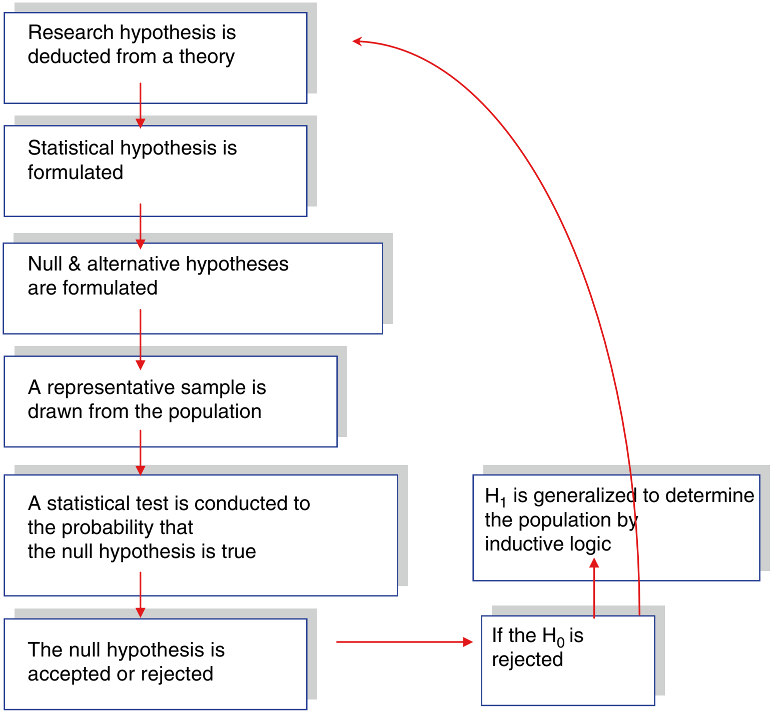

Research hypothesis is deducted from a theory
Statistical hypothesis is formulated
Null & alternative hypotheses are formulated
A representative sample is drawn from the population
A statistical test is conducted to the probability that the null hypothesis is true
The null hypothesis is accepted or rejected
If the H_0 is rejected
H_1 is generalized to determine the population by inductive logic

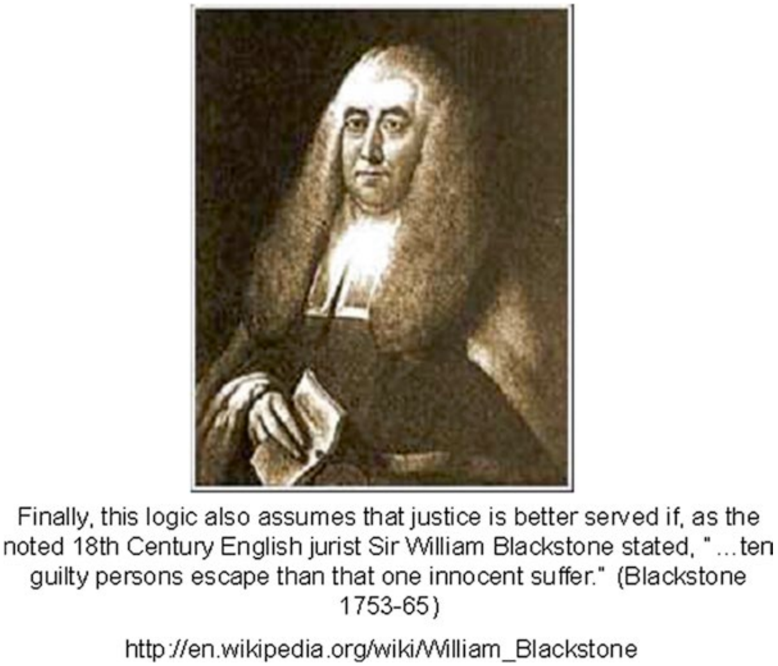
Fig. 2.3 Deductive and inductive logic of hypothesis testing

In the Anglo-American legal tradition, the consequences of these two possible errors are not considered equivalent. On the contrary, considerable safeguards have been incorporated into the criminal law to minimize the probability (α) of making a Type I error (convicting an innocent person), even at the risk of increasing the probability (β) of making a Type II error (releasing a guilty person). Indeed, this is where the concept of innocent until proven guilty comes from, and the quote: Finally, this logic also assumes that justice is better served if, as the noted 18th Century English jurist Sir William Blackstone stated, "…ten guilty persons escape than that one innocent suffer" [4, p. 358] (Fig. 2.4).

It is logical and critical to distinguish between the concepts of not guilty and innocent in the decision paradigm used in criminal law, i.e.:

If H_A = guilty, then does …
H_0 = not guilty, or does …
H_0 = innocent?

Here, "not guilty" does not mean the same thing as innocent. A not guilty verdict means that the evidence failed to convince the jury of the defendant's guilt beyond a reasonable doubt (i.e. "The scientific corollary is that data in this study was



Finally, this logic also assumes that justice is better served if, as the noted 18th Century English jurist Sir William Blackstone stated, "...ten guilty persons escape than that one innocent suffer." (Blackstone 1753-65)

http://en.wikipedia.org/wiki/William_Blackstone

Fig. 2.4 Sir William Blackstone quote regarding guilt and innocence

insufficient to determine if a difference exists, rather than there is no difference"). By this logic it is quite conceivable that a defendant can be found legally not guilty and yet not be innocent of having committed the crime in question.

The evaluation of a hypothesis involves both deductive and inductive logic. The process both begins and ends with the research hypothesis.

Step 1 Beginning with a theory about the phenomenon of interest, a research hypothesis is deduced.

This hypothesis is then refined into a statistical hypothesis about the parameters in the population.

The statistical hypothesis may concern population means, variances, medians, correlations, proportions, or other statistical measures.

The statistical hypothesis is then reduced to two mutually exclusive and collectively exhaustive hypotheses that are called the null (H_0) and alternative hypothesis (H_A).

Step 2 If the population is too large to study in its entirety (the usual case), a representative sample is drawn from the population with the expectation that the sample statistics will be representative of the population parameters of interest.

Step 3 The data gathered on the sample are subjected to an appropriate statistical test to determine if the sample with its statistical characteristics could have come from the associated population if the null hypothesis is true.

Step 4 Assuming that the null hypothesis (H_0) is true in the population, and that the probability that the sample came from such a population is very small ($p \leq 0.05$), the null hypothesis is rejected.

Step 5 Having rejected the null hypothesis, the alternative hypothesis (H_A) is accepted, and, by inductive inference is generalized to the population from whence the sample came.

These five steps are illustrated in Fig. 2.3, that is, the conduct of research involves a progressive generation of four kinds of hypotheses: Research hypothesis, Statistical hypothesis, Null hypothesis; and, Alternative hypothesis.

A research hypothesis is an affirmative statement about the relationship between two variables. For instance, consider the following example of a research hypothesis: “there is a positive correlation between the level of educational achievement of citizens and their support of rehabilitation programs for criminal offenders”. From the research hypotheses three other kinds of hypotheses can be formulated:

A statistical hypothesis
A null hypothesis
An alternative hypothesis

Again, a statistical hypothesis is a statement about the parameters of a population. The null hypothesis, which is symbolized H_0, is the negative statement of the statistical hypothesis; and, the alternative hypothesis, usually symbolized H_A, is the obverse of the null hypothesis and by custom, is stated to correspond to the research hypothesis being tested. Statements that are mutually exclusive are such that one or the other statement must be true. They cannot both be true at the same time. For instance:

Something is either “A” or “not A”. It cannot be both “A” and “not A” at the same time.
Or, the object on the kitchen table is either an apple or a non-apple.
Saying the object on the kitchen table is either an “apple” or a “non-apple” covers every possible thing that the object could be.

It is critical to understand that it is the null hypothesis (H_0) that is actually tested when the data are statistically analyzed, not the alternative hypothesis (H_A). Since H_0 and H_A are mutually exclusive, if the analysis of the data leads to the rejection of the null hypothesis (H_0), the only tenable alternative is to accept the alternative hypothesis (H_A). But, this does not mean that the alternative hypothesis is true, it may or may not be true. When we reject the null hypothesis it is because there is only a remote possibility that the sample could have come from a population in which the null hypothesis is true. Could we be wrong? Yes, and that probability is called alpha (α), and the error associated with alpha is called a Type I error (Table 2.2).

What about the converse situation, accepting the null hypothesis? If the null hypothesis is accepted, the alternative hypothesis may or may not be false. For example, the null hypothesis may be accepted because the sample size was too small to achieve the required degrees of freedom for statistical significance; or, an uncontrolled

Table 2.2 Compares the US legal system determination of guilt and innocence, to the Ho and Ha

	The verdict	The verdict
The truth	Accused in not guilty	Accused is guilty
Accused is not guilty (Ho true)	Justice is served	Innocent person is convicted probability = α
Accused is guilty (Ho false)	Guilty person is set free probability = β	Justice is served

extraneous variable or spurious variable has masked the true relationship between the variables; or, that the measures of the variables involved are grossly unreliable, etc. The issue is the same as a "not guilty" verdict in a criminal trial. That is, a verdict of not guilty does not necessarily mean that the defendant is innocent, it only means that the evidence was not sufficient enough to establish guilt beyond a reasonable doubt. There is a further discussion about the null hypothesis in Chap. 18.

An Overview of the Common Clinical Research Designs (Tables 2.3 and 2.4)

The common clinical research designs are listed in Tables 2.3 and 2.4 and summarizes some of their characteristics. There are many ways to classify study designs but two general ways are to separate them into descriptive and analytic studies and observational and experimental studies. These designations are fairly straightforward. In descriptive studies one characterizes (describes) a group of subjects; for example 'we describe the characteristics of 100 subjects taking prophylactic aspirin in the stroke belt.' In contrast, with analytic studies where there is a comparator group, for example, 'we compared the characteristics of 100 subjects in the stroke belt taking aspirin to 100 subjects not taking aspirin'. In experimental studies the investigator is "controlling" the intervention in contrast to observational studies where the exposure (intervention) of interest is occurring in nature and as the investigator you are observing the subjects with and without the exposure. Basically, experimental trials are clinical trials, and if subjects are randomized into the intervention and control (comparator) group it is a RCT.

Ecologic Studies

An ecological study is an epidemiological study in which the unit of analysis is a population rather than an individual. Ecologic studies are usually regarded as inferior to non-ecological designs such as cohort and case-control studies because of ecological fallacy (ecological fallacy refers to when inferences about the nature of individuals are deduced from inference for the population to which those individuals belong).

Table 2.3 General overview of study types

Observational
Ecologic studies
Case reports
Case series
Cross-sectional
Case-control
Cohort
Experimental
Clinical trials
Group clinical trials

Table 2.4 Types and descriptions of observational trials

Descriptive	Definition	Best used	Limitations
Case-series	Describes clinical course of one or more patients	Identify pathological, disease or treatment patterns	No comparison group
Ecologic	Associations of exposures and outcomes over time extracted from large databases	Trends over time	Impossible to adjust for confounding; Ecologic fallacy
Cross-sectional	Associations in a population at a single point in time	Generate data for further study	No temporality

Ecological studies can be easily confused with cohort studies, especially if different cohorts are located in different places. The difference is that in the case of ecological studies there is no information available about the individual members of the populations compared; whereas in a cohort study the data pair exposure/health is known for each individual. Ecologic studies use available population data to determine associations. For example, to determine an association between coronary heart disease (CHD) and the intake of saturated fat, one could access public records of beef sales in different states (or counties or regions of the country) and determine if an association existed between sales and the prevalence of CHD. Another example is that one might look for geographical correlations between disease incidence or mortality and the prevalence of risk factors. For example, mortality from coronary heart disease in local authority areas of England and Wales has been correlated with neonatal mortality in the same places 70 and more years earlier. This observation generated the hypothesis that coronary heart disease may result from the impaired development of blood vessels and other tissues in fetal life and infancy.

Case Reports and Case Series

Case reports and case series are potential ways to suggest an association, but, although limited in this regard, should not be deemed unimportant. For example, the recognition of the association of the diet drug combination of Fen-phen was the result of a case series [5]. These are, for the most part, descriptive observations about a single patient or a group of patients relative to some outcome of interest. It can be retrospective or prospective and usually involves a smaller number of patients than more powerful case-control studies or randomized controlled trials. Case series may be *consecutive* or *non-consecutive,* depending on whether all cases presenting to the reporting authors over a period were included, or only a selection. Case series may be confounded by selection bias, which limits statements on the causality of correlations observed; for example, physicians who look at patients with a certain illness and a suspected linked exposure will have a selection bias in that they have drawn their patients from a narrow selection (namely their hospital or clinic).

Cross-Sectional Studies

In cross-sectional studies, one measures and/or describes disease status (or outcome), exposure(s), and other characteristics at a point in time (point in time is the operative phrase), in order to evaluate associations between them. Cross-sectional studies are different from cohort studies in that cohort studies observe the association between a naturally occurring exposure and outcome (e.g., between health and a disease or between disease and an event) over a period of time rather than at a point in time. With cross-sectional studies, the exposure and outcome are evaluated at a point in time-i.e. there is no follow-up period where a subsequent evaluation of exposure/outcome is observed. Indeed, this measure “at a point in time” is both the strength and weakness of the cross-sectional (X-sectional) study design. Lack of a follow-up period means the study can be performed more rapidly and less expensively than a cohort study, but one sacrifices temporality (an important component for determining causality). In addition, because cross-sectional studies are evaluating cases (disease, outcomes) at a point in time, one is dealing with prevalent cases (not incident cases as is true of a cohort study). Confusing to some is that a X-Sectional study may take years to complete, so it is not the duration of the study that determines whether it is a x-sectional or a cohort design, it is the time between the exposure and outcome that makes that determination. In other words, if the exposure and outcome are measured at a single point in time, it is x-sectional. If the outcome is ascertained at some time point distant from the exposure it is a cohort study. There are a number of factors that must be considered when using prevalence (rather than incidence) and these are summarized in Fig. 2.5.

An example of a cross sectional study might be the assessment of arterial stiffness and hormone replacement therapy (HRT). Let’s say a study is designed where age matched women receiving HRT are compared to women not taking HRT, and arterial stiffness is measured in each to determine if differences in arterial stiffness differ between the two groups. Some have likened this to taking a snapshot of the association at that point in time.

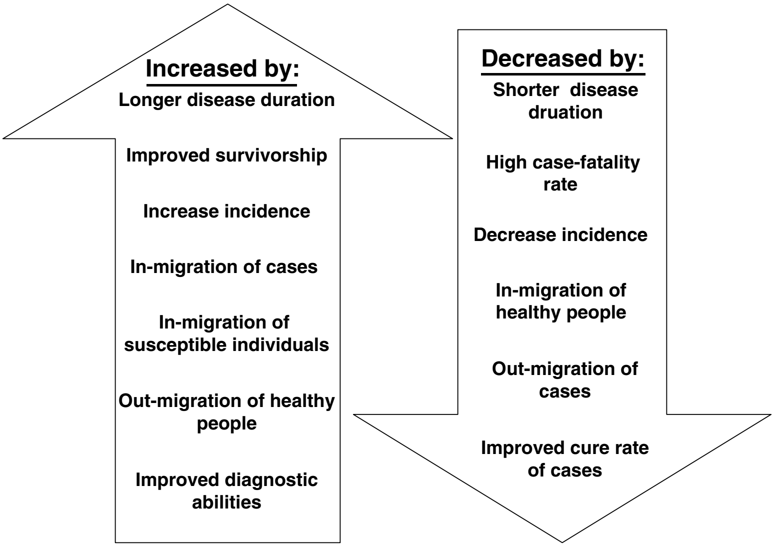


Fig. 2.5 The balance of factors that affect prevalence

Case-Control Study

In a case-control study (CCS), the investigator identifies a certain outcome in the population, then matches the 'diseased group' to a 'healthy group,' and finally identifies differences in exposures between the two groups.

With a CCS one approaches the study design the opposite of a cohort design (in fact some have suggested the use of the term 'trohoc design' – cohort spelled backwards). The term case-control study was coined by Sartwell to overcome the implication that the retrospective nature of the design was an essential feature [6]. That is, patients with the outcome of interest are identified, a control group is selected, and one then looks back for exposures that differ between the two. Two major biases exist with the CCS; first the selection of the control group is problematic, and second, one is usually looking back in time (i.e. it is a retrospective study in that sense). Selecting the control group for a CCS is problematic because if one selects too many matching criteria it becomes difficult to find an adequate control group, while if one has too few matching criteria, the two groups can differ in important variables. For CCS designs, recall bias is also an issue (this is even a greater issue if death is an outcome, in which case one not only has to deal with recall bias, but the recall is obtained from family members, caregivers, etc. rather than the subject).

One of the strengths of the CCS design is that if one is interested in a rare disease, one can search the area for those cases, in contrast to randomly selecting a cohort population that will develop this rare disease infrequently, even over a long follow-up time period. Also, in contrast to a cohort study in which the sample population is followed for a time period, a CCS obviates this need so one can complete the study much more rapidly (and therefore less expensively).

There are several variations of the case-control design that overcome some of the shortcomings of a typical CCS (although they have their own limitations): a prospective CCS and a nested CCS. In the prospective CCS, one accrues the cases over time (i.e. in a prospective fashion) so that recall bias is less of an issue. However, one then has to wait until enough cases are accrued (problematic again for rare diseases); and, the selection of an appropriate control group still exists. A nested case-control study is a type of study design where outcomes that occurred during the course of a cohort study or RCT are compared to controls selected from the same cohort or clinical trial population who did not have the outcome. Compared with the typical case-control study, a nested case-control study can reduce 'recall bias' and temporal ambiguity, and compared with a cohort study, it can reduce cost and save time. One additional drawback of a nested case-control study is that the non-diseased persons from whom the controls are selected may not be fully representative of the original cohort as a result of death or failure to follow-up cases. As mentioned, the nested CCS design can be placed within a cohort study or RCT. An example is taken from the Cholesterol and Recurrent Events (CARE) Study [7]. The primary study was aimed at the prevention of recurrent MI when patients with a prior MI and 'normal' cholesterol levels were further treated with pravastatin. As part of the original study plasma was stored and after the report of the primary study was published the following study was designed: "we conducted a prospective, nested case-control study in the Cholesterol and Recurrent Events (CARE) trial. Baseline concentrations of VLDL-apolipoprotein (apo) B (the VLDL particle concentration), VLDL lipids, and apoCIII and apoE in VLDL+LDL and in HDL were compared in patients who had either a myocardial infarction or coronary death (cases, n=418) with those in patients who did not have a cardiovascular event (control subjects, n=370) in 5 years of follow-up. VLDL-cholesterol, VLDL-triglyceride, VLDL-apoB, apoCIII and apoE in VLDL+LDL and apoE in HDL were all interrelated, and each was a univariate predictor of subsequent coronary events. Adjustment for LDL- and HDL-cholesterol did not affect these results" [7].

Cohort Study

A cohort study is much like a RCT except that the intervention in an RCT is "investigator controlled", while in a cohort study the intervention (exposure) is a naturally occurring phenomenon. A cohort design is a study in which two or more groups of people that are "free of disease" at study onset and that differ according to the extent of exposure (e.g. exposed and unexposed) are compared with respect to disease incidence. A cohort study assembles a group of subjects and follows them over time. One follows these subjects to the development of an outcome of interest and then compares the characteristics of the subjects with and without the outcome in order to identify risk factors (exposures) for that outcome. A major assumption

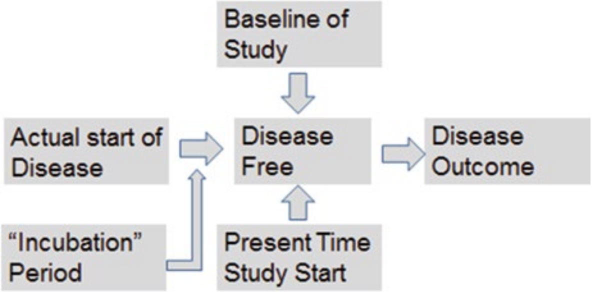


Fig. 2.6 The cohort limitation

made in cohort studies is that the subject is disease free at the beginning of the study (disease free means for the outcome- disease- of interest). For example, if the outcome of interest is a ***recurrent*** myocardial infarction, the subject would have had the first infarction (so in that sense he is not disease free) but in terms of the outcome of interest (a second infarction) we assume that at study onset, he is not having a second infarction. This example may seem obvious, but let us use colon cancer as another example. At study onset, one assumes that the subject is disease free (cancer-free or 'normal') at the time of enrollment, while in fact he or she may already have colon cancer that is as yet undiagnosed. This could bias the results of the study since the exposure of interest may have nothing to do with the outcome of interest (colon cancer) since the subject already has the outcome irrespective of the exposure (say a high fat diet). This also raises the issue as to what is 'normal'. One whit suggested that a normal subject is one that has been insufficiently tested! The cohort assumption mentioned above is diagrammed in Fig. 2.6. Of course, one also assumes that the incorrect assumption of no disease at onset is equally balanced in the two groups under study, and that is indeed the hope, but not always the realization. Cohort studies are considered the best way to study prognosis, but one can also do this by using a case-control design.

As an example, recall the cross-sectional study described above (the example of a cross sectional study that assessed the association of arterial stiffness and hormone replacement therapy). Let's say a study is designed where age matched women receiving HRT are compared to women not taking HRT, and arterial stiffness is measured in each to determine if differences in arterial stiffness differ between the two groups. Suppose now we follow subjects for 5 years, measure their arterial stiffness, and determine if there is a difference in that measure in women receiving HRT compared to those who are not. This would be a cohort design.

Retrospective Cohort Design

Cohort studies are generally prospective; however, retrospective cohort studies do exist. The key to the study design is identifying the exposure of interest in 'normal' subjects without disease (i.e. the outcome of interest), evaluate for that outcome

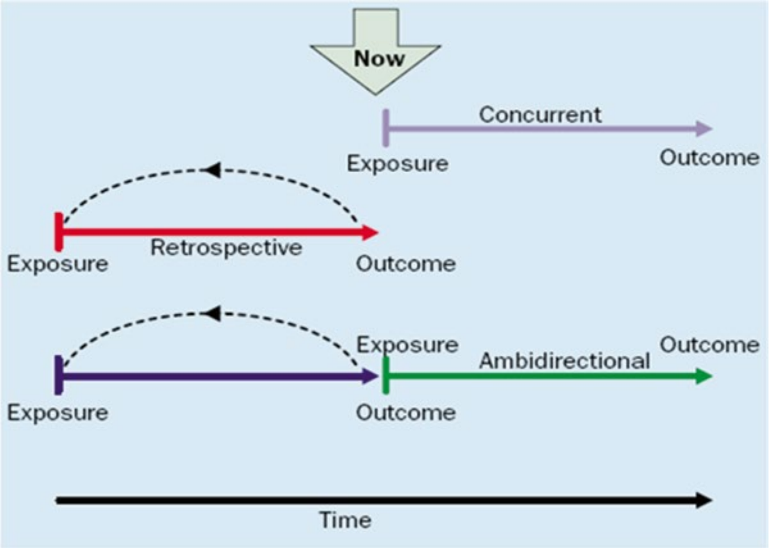


Fig. 2.7 A comparison of prospective and retrospective cohort study designs

after a period of time has elapsed, and determining the exposure as different or not in those with and without the outcome. Retrospective cohort studies are particularly well suited to the study of long-term occupational hazards. An example of a retrospective cohort study is the study of nickel refinery workers where about 1,000 nickel refinery workers were identified from company records and their outcomes identified over a prior 10 year period. Sixteen were found to have died from lung cancer (expected rate was 1 from National data), 11 died from nasal cancer (1 expected) and 67 from other causes (72 expected) [8].

Or, to continue with our HRT example from above, suppose we now identify a group of women who have and have not been taking HRT and we now measure their arterial stiffness and make comparisons of association to HRT (Fig. 2.7). It is common that this design is confused with case control studies. The differentiating factor is whether one designs the study to evaluate whether the exposure (e.g. HRT in this example) is associated with the outcome (arterial stiffness), this would be a cohort design; or, if subjects are identified by their outcome (say normal vs. abnormal arterial stiffness) and then exposure status is determined (they did or did not take HRT), this would be a case control design.

Case Cohort Design

Another modification of cohort studies is the case-cohort design. With the case-cohort design, a ‘subcohort’ is randomly selected from the cohort sample, a separate exposure of interest from the total cohort is identified, and cases (outcomes) are then determined in the same manner as the primary design. An example might be a cohort study of 10,000 subjects that is assessing some outcome-let's say a CVD outcome- in relation to dietary fat. The investigator decides that she would also like to know the association of CVD with a measure of coronary artery calcium, so electron beam computed tomography (EBCT-a relatively expensive procedure to perform on the all of the original cohort) is measured in a random sample of 100 of the cohort subjects (the 'subcohort'). The association of EBCT to CVD outcome is then ultimately determined. A key feature of this design is that the cases are selected from among those with disease, while the controls are selected at the beginning of the study period, irrespective of disease status (that is some control cases may later become a case).

Randomized Control Trial (RCT)

In the randomized-controlled trial (RCT), the exposure is "controlled" by the investigator, which contrasts it to all the other study designs. A detailed discussion of the RCT will be presented in Chap. 3. However, it should be noted that RCTs cannot be used to address all important questions. For example, observational studies are more appropriate when studies are used to detect rare or late consequences of interventions, situations not best suited to the RCT.

The above discussion of study designs is not meant to be all-inclusive. For example there is a design called the "case-only design" that is somewhat unique to genetic studies. The case-only design is an efficient and valid approach to screening for gene-environment interaction under the assumption of the independence between exposure and genotype in the population. That is, if the primary purpose of the study is to estimate the effect of gene-environment interaction in disease etiology, one can do so without employing controls, thus, the case-only design requires fewer cases than the case-control design to measure gene-environment interaction, and it also requires fewer cases to measure gene-gene interactions.

One should now be able to begin to understand the key differences, and therefore limitations, of each study design; and, circumstances where one design might be preferable to another. Let's, for example, use the exposure of electromagnetic energy (EME) and cancer outcome (e.g. leukemia). With a cross-sectional study, a population is identified (target population), cancer rates determined, and exposure and lack of exposure to EME is ascertained from a sample population. One then analyzes the exposure rates in subjects with cancer and those that are cancer free. If the cancer rate is higher in those who were exposed, an association is implied. This would be a relatively inexpensive way to begin to look at the possible association of these variables, but limitations should be obvious. For example, since there is no temporality in this type of design, and since biologically, exposure to EME if it did cause cancer would likely have to occur over a long period of time, one could easily miss an association. Also reverse causation cannot be ruled out. Also remember, that even though the RCT is generally the "best"

Table 2.5 Common study designs, uses and limitations

Descriptive	Definition	Best used	Limitations
Cohort	Comparison of outcome of those with and without exposure	Rare exposure, common outcomes	Lack of randomization, bias from dropouts
Case-cohort	Exposure between cases and random sample of the original cohort	Rare exposure and outcome, long latency period	Recall bias; not suitable for chronic conditions
Case-crossover	Each case contributes one case window of time and one or more control windows	Outcome does not vary over time; exposures are brief	Recall bias; not suitable for chronic conditions
Cross-sectional	Description of associations at a single point in time	Outcome associations to generate further study	No temporality
Case-control	Odds of exposure among cases c/w non-cases	Common exposure rare outcome	Selection and recall bias; confounding
Nested case-control	Case-control nested within cohort (or clinical trial)	Rare outcome and/or long latency period	Decreases biases of case-control Design

study design, one could easily see why it would not be appropriate for this research question. Table 2.5 summarizes a few of the study designs in relation to the frequency of the exposure and outcome.

In summary, it should be evident that observational studies (e.g. cross-sectional, case-control, and cohort studies) have a major role in research. However, despite their important role, von Elm et al. discussed the lack of important information that was either missing or unclear in prior published observational studies; and why this lack of information led to a guideline document for reporting observational studies (the STROBE statement-the Strengthening and Reporting of Observational Studies in Epidemiology). The STROBE statement was designed after the CONSORT- the Consolidated Standards of Reporting Trials-; this statement outlines the guidelines for reporting RCTs. The STROBE statement is a checklist of 22 items that are to be considered essential for good reporting of observational studies (also see Chap. 19) [9].

Formulating relevant and precise questions that can be answered can be complex and time consuming. A structured approach for framing questions that uses five components may help facilitate the process. This approach is commonly known by the acronym "PICOS" or "PECOS", where each letter refers to a component as follows:

- P refers to the patient population or the disease being addressed,
- I (or E) refers to the intervention or exposure
- C refers to the comparator group
- O to the outcome or endpoint
- S refers to the study design chosen

Finally, the spectrum of evidence imparted by the different clinical research designs ranges from ecological studies through observational epidemiological studies to randomized control trials (RCTs). And, many people are becoming increasingly skeptical of RCTs. In fact, one researcher has claimed that 90 % of medical research is wrong [10]. Examples include: Two 1993 studies concluded that vitamin E prevents cardiovascular disease. That claim was overturned in 1996 and 2000; a 1996 study concluding that estrogen therapy reduces older women's risk of Alzheimer's was overturned in 2004; and, a major study concluded there's no evidence that statins help people with no history of heart disease –the cost of statins is more than $20 billion per year, of which half may be unnecessary. Jeffry Hyman has reviewed this subject and has published an On-Line tutorial that addresses this question [11]. Hyman points out the following in his presentation entitled "Is Most Medical Research Wrong? The Role Of Incentives And Statistical Significance: a myriad of biases are present in any type of research that includes selection bias, information bias (see Chap. 17) and a number of analytical issues (see Chap. 3). In addition, Hyman points out the power of incentives by raising the questions of whether we are looking for the truth when we do research… or are we? and is the search for truth our only reason for doing research? could we have any other incentives? In answer to this latter question he raises the financial and egocentric motivations for doing research beyond seeking the truth, such as:

- We want to get grants
- We have a financial interest in the study
- We might want to support funding for a program
- We might want to continue funding for a program
- We want tenure
- We want a promotion
- We think there is an association and we want to show it
- We want our studies to be published
- We want publicity
- We might have done work in this area before and we want to replicate it
- We have made an Investment of time and money in doing the study
- we want to show results
- We want people to think we are a good researcher
- We know about publication bias towards negative results

He asks an additional question: how does our strong desire for a $P<0.05$ affect our results? and points to the following "follies": We do extensive modeling with a range of variables. Then we only report the model with the most significant results (selective reporting) (multiple comparisons), we do extensive subgroup analyses (multiple comparisons), we compare extreme groups, such as the 1st and 5th quintiles, we use too large of a sample size for the effect we want to measure (such as national surveys), we use a 1 sided P value, we don't try to publish papers with negative results, we quickly do studies in hot fields, we change study endpoints after looking at the data,

Table 2.6 Study designs by frequency of exposure and outcomes

	Prevalence or incidence of outcome	
Drug exposure	Not rare	Rare
Not rare	Cohort or clinical trial	Case-control
Rare	Cohort	Case-control

we investigate multiple associations between exposure and outcome, and, we selectively site the literature. A hypothetical (extreme) example of these latter concepts is presented in Hymans presentation (in whom he cites www.johndcook.com) follows: Researchers test 200 completely ineffective new drugs;

- About 10 trials out of the 200 will have a "significant" result due to chance.
- Only the 10 studies with significant results will be submitted for publication.
- Five of these studies are published in major journals
- Result: The type 1 error rate of each study was 5 %, but the error rate in the literature is 100 %

Hyman concludes with the question "Can we predict which studies are more likely to be wrong?". Here is a list of his answers: small studies with significant results; studies with more flexible designs, outcome measures, and models; studies with significant results and a small effect measure (like odds = 1.1); The hotter the field and the more people doing research in it; studies where there are strong financial interests; studies with strong pre-existing beliefs by researchers. In summary: be aware of how we overstate our results in an effort to get statistically significant results; be aware of the limitations of P values and statistical significance; don't over interpret significant results, being significant does not make a results true or important; on the other hand, not being significant does not make a result false; do power calculations for each study and don't make your study too big or too small, make it just right; watch for problems like multiple comparisons, subgroup analysis, and selective reporting; be aware of the situations where study results are more likely to be wrong; remember the effects of publication bias; and, in observational studies speak about associations, not causality (Table 2.6).

Finally, it has been pointed out by some, that clinical trials are too expensive, recruit too few patients, and results in to many investigators to just give up because of the cost and complexity of clinical trials (in fact it was noted that 38 % of PIs who participated in clinical trials between 2000 and 2005, did not return to conduct another clinical trial [12]). It has also been suggested that half of RCTs never finish due to recruitment problems, and many that do finish are underpowered to answer the original research question, even as costs soar. As a solution, it has been suggested that since observational trials give results similar to RCTs, and at less expense, they can be used as a substitute [13]. While Pocock and Elbourne warn that the one critical deficiency of observational designs is the absence of the randomization that occurs with RCTs rather than each patients treatment being deliberately chosen in observational trials [14].

References

1. Parker Palmer. Accessed at http://en.wikipedia.org/wiki/Parker_Palmer
2. Vickers AJ. Michael Jordan won't accept the null hypothesis: notes on interpreting high P values. Medscape. 2006;7:3.
3. The Null Logic of Hypothesis Testing. Accessed at http://www.shsu.edu/~icc_cmf/cj_787/research6.doc
4. Cited in The Null Logic of Hypothesis Testing. 1753–65. Accessed at http://www.shsu.edu/~icc_cmf/cj_787/research6.doc
5. Connolly HM, Crary JL, McGoon MD, Hensrud DD, Edwards BS, Edwards WD, et al. Valvular heart disease associated with fenfluramine-phentermine. N Engl J Med. 1997;337:581–8.
6. Cited in Sartwell P and Nathanson N. Epidemiol Rev; 1993.
7. Sacks FM, Pfeffer MA, Moye LA, Rouleau JL, Rutherford JD, Cole TG, et al. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. Cholesterol and recurrent events trial investigators. N Engl J Med. 1996;335:1001–9.
8. Doll R. Cohort studies: history of the method II. Retrospective cohort studies. Soz Praventivmed. 2001;46:152–60.
9. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147:573–7.
10. Freedman D. A researchers claim: 90% of medical research is wrong. Health Fam. 2010;2:1–2.
11. Hyman J. Why most published research findings are false. Am J Epidemiol. 2005;2:e124.
12. CardioSource World News. Nov 2013:25–35.
13. Benson K, Hartz AJ. A comparison of observational studies and randomized controlled trials. N Engl J Med. 2000;342:1878–86.
14. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations. N Engl J Med. 2000;342:1907–9.

Chapter 3
A Focus on Clinical Trials

Stephen P. Glasser

A researcher is in a gondola of a balloon that loses lift and lands in the middle of a field near a road. Of course, it looks like the balloon landed in the middle of nowhere. As the researcher ponders appropriate courses of action, another person wanders by. The researcher asks, 'Where am I?' The other person responds, 'You are in the gondola of a balloon in the middle of a field.' The researcher comments, 'You must design clinical trials.' 'Well, that's amazing, how did you know?' 'Your answer was correct and precise and totally useless.' (ANON)

Abstract The spectrum of evidence imparted by the different clinical research designs ranges from ecological studies through observational epidemiological studies to randomized control trials (RCTs). This chapter addresses the definition of clinical research, the major aspects of clinical trials e.g. ethics, randomization, masking, recruitment and retention of subjects enrolled in a clinical trial, patients/subjects lost to follow-up during the trial etc. Although this chapter focuses on the weaknesses of clinical trials, it is emphasized that the randomized, placebo-controlled, double blind clinical trial is the design that yields the greatest level of scientific evidence.

Keywords Generalizability/external validity. Internal validity • Superiority testing • Equivalence/noninferiority testing • Randomization • Intention to treat • Missing

Less than 50 % of this chapter is taken from "Clinical trial design issues: at least 10 things you should look for in clinical trials" [1] with permission of the publisher.

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_3,
© Springer International Publishing Switzerland 2014

data • Eligibility • Efficacy/effectiveness • Blinding/masking • Subgroup analysis • Surrogate endpoints • Composite endpoints • Primary and secondary endpoints

The differences in clinical research designs and the different weights of evidence imparted by different clinical research designs, are exemplified by the post-menopausal hormone replacement therapy (HRT) controversy. Multiple observational epidemiological studies had shown that HRT was strongly associated with the reduction of atherosclerosis, myocardial infarction risk, and stroke risk [2–4]. Subsequently, 3 RCTs suggested that HRT was not beneficial, and might even be harmful [5–7]. This latter observation raises a number of questions, including: why can this paradox occur? What can contribute to this disagreement?; and, why do we believe these 3 RCT's more than so many well-done observational trials? The reasons for this are many (also see Chap. 2), but include: concerns about the generalizability of clinical trial results to the general population, and the reproducibility of the results; and, RCTs are increasingly involving thousands of patients form many sites, and from multiple countries making them challenging to design and difficult to execute and monitor [8]. Also, some clinical trials have been criticized by regulatory agencies due to apparent high dropout rates and patients lost to follow up, which has led to new FDA guidelines emphasizing the importance of patient retention and innovative site monitoring [9]. In support of this latter issue, is a post hoc analysis of the Efficacy of Vasopressin Antagonism in Heart Failure: Outcome Study with Tolvaptin (EVEREST) in which the authors evaluated the relationship between the number of patients enrolled in each site with trial outcomes. They found that the high enrolling sites had better clinical outcomes and more protocol completion rates compared to the lower enrolling sites [10]. Of course, there are a number of explanations for this observation from EVEREST, and as was pointed out in the discussion of this trial, the use of block randomization (see below) within each center should have equally distributed patients between the sites of potentially differing quality who were on or off study drug; none-the-less, the point is one worthy of further research [10]. Participant differences based on geographic disparities have been well described, but differences related to participant volume have not.

Frequently, there is confusion about the difference between clinical research and clinical trials. In general usage experimental design is the design of any information-gathering exercises where variation is present, whether under the full control of the experimenter or not. Other types of study are opinion polls and statistical surveys (which are types of observational study), natural experiments and quasi-experiments. In the design of experiments, the experimenter is often interested in the effect of some process or intervention (the "treatment") on some objects (the "experimental units"), which may be people, parts of people, groups of people, plants, animals, materials, etc.

A clinical trial is a type of experimental study undertaken to assess the response of an individual (or in the case of group clinical trials-a population) to interventions introduced by an investigator. Clinical trials can be randomized or non-randomized, un-blinded, single-blinded, or double-blinded; comparator groups can be placebo, active controls, or no treatment controls, and RCTs can have a variety of designs (e.g. parallel group, crossover, etc.). That being said, the RCT remains the 'gold-standard' study design and its results are appropriately credited as yielding the highest level of scientific evidence (greatest likelihood of causation). However, recognition of the limitations of the RCT is also important so that results from RCTs are not blindly accepted. As Grimes and Schultz point out, in this era of increasing demands on a clinician's time it is 'difficult to stay abreast of the literature, much less read it critically. In our view, this has led to the somewhat uncritical acceptance of the results of a randomized clinical trial' [11]. Also, Loscalzo, has pointed out that 'errors in clinical trial design and statistical assessment are, unfortunately, more common that a careful student of the art should accept' [12].

What leads the RCT to the highest level of evidence and what are the features of the RCT that renders it so useful? Arguably, one of the most important issues in clinical trials is having matched groups in the interventional and control arms; and, this is best accomplished by randomization. That is, to the degree that the two groups under study are different, results can be confounded by any difference, while when the two groups are similar, confounding is reduced (see Chap. 17 for a discussion of confounding). It is true that when potential confounding variables are known, one can relatively easily adjust for them in the design or analysis phase of the study. For example, if one believes that smoking might confound the results of the success of treatment for hypertension, one can build into the design a stratification scheme that separates smokers form non-smokers, before the intervention is administered and in that way determine if there are differential effects in the success of treatment (e.g. smokers and non-smokers are randomized equally to the intervention and control). Conversely, one can adjust after data collection in the analysis phase by separating the smokers from the non-smokers and again analyze them separately in terms of the success of the intervention compared to the control. The real challenge of clinical research, is not how to adjust for **known** confounders, but how to have matched variables in the intervention and control arms, when potential confounders are **not** known. Optimal matching is accomplished with randomization, and this is why randomization is so important. More about randomization later, but in the meanwhile one can begin to ponder how un-matching might occur even in a RCT. In addition to randomization, there are a number of important considerations that exist regarding the conduct of a clinical trial, such as: is it ethical? What type of comparator group should be used? What type of design and analysis technique will be utilized? How many subjects are needed and how will they be recruited and retained?

Finally, there are issues unique to RCTs (e.g. intention-to-treat analysis, placebo control groups, randomization, equivalence testing) and issues common to all clinical research (e.g. ethical issues, blinding, selection of the control group, choice of the outcome/endpoint, trial duration, etc.) that must be considered (Table 3.1). Each of these issues will be reviewed in this chapter. To this end, both the positive and problematic areas of RCTs will be highlighted.

Table 3.1 Issues of importance for RCTs

- Ethical considerations
- Randomization
- Eligibility criteria
- Efficacy vs. effectiveness
- Compliance
    - Run-in periods
    - Recruitment and retention
- Masking
- Comparison groups
    - Placebo
    - ‘Normals’
- Analytical issues
    - ITT
    - Subgroup analysis
    - Losses to follow-up
    - Equivalence vs. traditional testing
- Outcome selection
    - Surrogate endpoints
    - Composite endpoints
    - Trial duration
- Interpretation of results
- Causal inference
- The media role in reporting RCT results

Ethical Issues

Consideration of ethical issues is key to the selection of the study design chosen for a given research question/hypothesis. For RCTs ethical considerations can be particularly problematic, mostly (but by no means solely) as it relates to using a placebo control. A full discussion of the ethics of clinical research is beyond the scope of this book, and for further discussion one should review the references noted here [13–15]. (There is also further discussion of this issue under the section entitled “Traditional vs. Equivalence Testing” and Chaps. 4 and 7). The opinions about when it is ethical to use placebo controls are quite broad. For example, Rothman and Michaels are of the opinion that the use of placebo is in direct violation of the Nuremberg Code and the Declaration of Helsinki [15], while others would argue that placebo controls are ethical as long as withholding effective treatment leads to no serious harm and if patients are fully informed. Most would agree that placebo is unethical if effective life-saving or life-prolonging therapy is available or if it is likely that the placebo group could suffer serious harm. For ailments that are not likely to be of harm or cause severe discomfort, some would argue that placebo is justifiable [14]. However, in the majority of scenarios, the use of a placebo control is not a clear-cut issue, and decisions need to be made on a case-by-case basis. One prevailing standard that provides a guideline for when to study an intervention against placebo is when one has enough confidence in the intervention that one is comfortable that the additional risk of exposing a subject to the intervention is low relative to no therapy or the 'standard' treatment; but, that there is sufficient doubt about the intervention that use of a placebo or active control ('standard treatment') is justified. This balance, commonly referred to as *equipoise*, can be difficult to come by and is likewise almost always controversial. Importantly, equipoise needs to be present not only for the field of study (i.e. there is agreement that there is not sufficient evidence of the superiority of an alternative treatments), but equipoise also has to be present for individual investigators (permitting individual investigators to ethically assign their patients to treatment at random).

Another development in the continued efforts to protect patient safety is the Data Safety and Monitoring Board (DSMB-see Chap. 9). The DSMB is now almost universally used in any long-term intervention trial. First a data and safety monitoring plan (DSMP) becomes part of the protocol, and then the DSMB meets at regular and at 'as needed' intervals during the study in order to address whether the study requires early discontinuation. As part of the DSMP, stopping rules for the RCT will have been delineated. Thus, if during the study, either the intervention or control group demonstrates a worsening outcome, or the intervention group is showing a clear benefit, or adverse events are greater in one group vs. the other (as defined within the DSMP) the DSMB can recommend that the study be stopped. But, the early stopping of studies can also be a problem. For example, in a recent systematic review by Montori et al., the question was posed about what was known regarding the epidemiology and reporting quality of RCTs involving interventions stopped for early benefit [16]. Their conclusions were that prematurely stopped RCTs often fail to adequately report relevant information about the decision to stop early, and that one should view the results of trials that are stopped early with skepticism [16].

Randomization

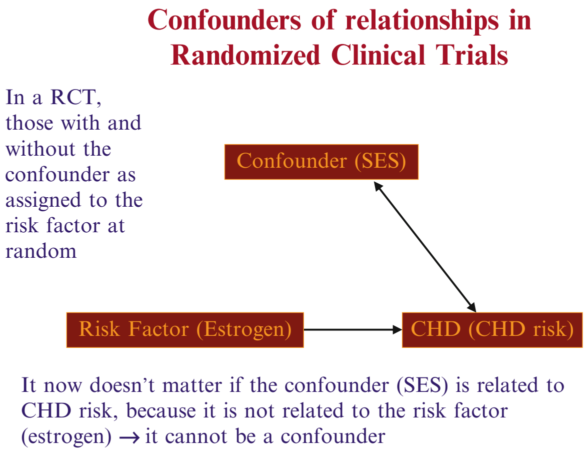
Arguably, it is randomization that results in the RCT yielding the highest level of scientific evidence (i.e. resulting in the greatest likelihood that the intervention is causally related to the outcome). Randomization is a method of treatment allocation that is a distribution of study subjects at random (i.e. by chance). As a result, randomization results in all randomized units (e.g. subjects) having the same and independent chance of being allocated to any of the treatment groups, and it is impossible to know in advance to which group a subject will be assigned. The introduction of randomization to clinical trials in the modern era can probably be credited to the 1948 trial of streptomycin for the treatment of tuberculosis [17]. In this trial, 55 patients were randomized to either streptomycin with bed rest, and were compared to treatment with bed rest alone (the standard treatment at that time). To quote from that paper, 'determination *of whether a patient would be treated by streptomycin and bed rest (S case) or bed rest alone (C case), was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each center by Professor Bradford Hill; the details of the series were unknown* *to any of the investigators or to the co-coordinator and were contained in a set of sealed envelopes each bearing on the outside only the name of the hospital and a number. After acceptance of a patient by the panel and before admission to the streptomycin centre, the appropriate numbered envelope was opened at the central office; the card inside told if the patient was to be an S or C cases, and this information was then given to the medical officer at the centre*'. Bradford Hill was later knighted for his contributions to science including the contribution of randomization.

With randomization the allocation ratio (number of units-subjects- randomized to the investigational arm versus the number randomized to the control arm) is usually 1:1. But a 1:1 ratio is not required, and there may be advantages to unequal allocation (e.g. 2:1 or even 3:1). The advantages of unequal allocation are: one exposes fewer patients to placebo, and one gains more information regarding the safety of the intervention. The main disadvantage of higher allocation ratios is the loss of power.

There are three general types of randomization: simple, blocked, and stratified. Simple randomization can be likened to the toss of an unbiased coin (i.e. heads group A, tails group B). This is easy to implement, but particularly with small sample sizes, could result in substantial imbalance (for example if one tosses a coin 10 times, it is not improbable that one could get 8 heads and 2 tails. If one tosses the coin 1,000 times it is likely that the distribution of heads to tails would be close to 500 heads and 500 tails). Blocked randomization (sometimes called permuted block randomization) is a technique common to multi-center studies. Whereas the entire trial might intend to enroll 1,000 patients, each center might only contribute 10 patients to the total. To prevent between center bias (recall each sample population has differences even if there is matching to known confounders) blocked randomization can be utilized. Blocked randomization means that randomization occurs within each center ensuring that about 5 patients in each center will be randomized to the intervention and 5 to the control. If this approach was not used, one center might enroll 10 patients to the intervention and another center, 10 patients to the control group. Recall that the main objective of randomization is to produce between-group comparability. If one knows prior to the study implementation that there might be differences that are not equally distributed between groups (again particularly more likely with small sample sizes) stratified randomization can be used. For example, if age might be an important indicator of drug efficacy, one can randomize within strata of age groups (e.g. 50–59, 60–69 etc.). Within each stratum, randomization can be simple or blocked.

In review, simple randomization is the individual allocation of subjects into the intervention and control groups, block randomization creates small groups (blocks) in which there are equal numbers in each treatment arm so that there are balanced numbers throughout a multi-center trial, and stratified randomization addresses the ability to separate known confounders into strata so that they can no longer confound the study results. Again, randomization is likely the most important key to valid study results because (if the sample size is large enough), it distributes known, and ***more importantly unknown,*** confounders equally to the intervention and control groups.

Fig. 3.1 The relationship of confounders to outcome and how they are eliminated in a RCT



Confounders of relationships in Randomized Clinical Trials

In a RCT, those with and without the confounder as assigned to the risk factor at random

It now doesn’t matter if the confounder (SES) is related to CHD risk, because it is not related to the risk factor (estrogen) → it cannot be a confounder

Now, as to the problems associated with randomization. As prior discussed, the issue of confounders of relationships is inherent in all clinical research. A confounder is a factor that is associated with both the risk factor and the outcome, and leads to a false apparent association between the risk factor and outcome (see Fig. 3.1). In observational studies, there are several approaches to remove the effect of confounders:

- Most commonly used in case/control studies, one can match the case and control populations on the levels of potential confounders. Through this matching the investigator is assured that both those with a positive outcome (cases) and a negative outcome (controls) have similar levels of the confounder. Since, by definition, a confounder has to be associated with both the risk factor and the outcome; and, since through matching the suspected confounder is not associated with the outcome – then the factor cannot affect the observed differences in the outcome. For example, in a study of stroke, one may match age and race for stroke cases and community controls, with the result that both those with and without strokes will have similar distributions for these variables, and differences in associations with other potential predictors are not likely to be confounded, for example, by higher rates in older or African American populations.
- In all types of observational epidemiological studies, one can statistically/mathematically ‘adjust’ for the confounders. Such an adjustment allows for the comparison between those with and without the risk factor at a ‘fixed level’ of the confounding factor. That is, the association between the exposure and the potential confounding factor is removed (those with and without the exposure are assessed at a common level of the confounder), and as such the potential confounder cannot bias the association between the exposure and the outcome. For example, in a longitudinal study assessing the potential impact of hypertension on stroke risk, the analysis can ‘adjust’ for race and other factors. This adjustment implies that those with and without the exposure (hypertension) are assessed as if race were not associated with both the exposure and outcome.

Table 3.2 Example of the use of propensity scoring

	Before matching			After matching		
Variable (%)	Aspirin	No aspirin	P value	Aspirin	No aspirin	P value
Men	77	56	<.001	70.4	72.1	.33
Diabetes	16.8	11.2	<.001	15	15.3	.83
HTN	53	40.6	<.001	50.3	51.7	.46
CAD Hx	69.7	20.1	<.001	48.3	48.8	.79
CHF	5.5	4.6	.12	5.8	6.6	.43
B-Blocker	35.1	14.2	<.001	26.1	26.5	.79
ACE I	13	11.4	<.001	15.5	15.8	.79

Adapted from: Gum et al. [19]

The Propensity Score has received increased interest. The propensity score was introduced by Rosenbaum and Rubin [18] to provide an alternative method for estimating treatment effects when treatment assignment can be assumed to be unconfounded but is not random. A propensity score is the probability of a unit (e.g., person, classroom, school) being assigned to a particular condition in a study given a set of known covariates (a variable that is possibly predictive of the outcome under study). In an attempt to simulate randomization, propensity scores are used to reduce selection bias by equating groups based upon covariates (this, balances known confounders, but obviously not the unknown confounders). In the analysis of treatment effects, suppose that we have a binary treatment T, an outcome Y, and background variables X. The propensity score is defined as the conditional probability of treatment given background variables. This is operationalized by gathering all the background information that we have on subjects before exposure is known and building a model to predict the probability that they will be in the exposed vs. unexposed group. Groups of subjects with similar propensity scores can then be expected in the aggregate to have similar values of all the background information. Thus, propensity scores can be used in cohort trials, clinical trials without randomization, administrative data base studies, detecting safety signals, secondary questions within RCTs; and, propensity score analyses may be used in either the design or analysis phase. One example of the use of the propensity score is the aspirin and mortality study reported by Gum et al. [19]. In that study, 6,174 subjects underwent stress echocardiography for the evaluation of known or suspected coronary artery disease. Aspirin was being taken by 37 % of the subjects. The main outcome was all cause mortality and the mean follow-up was 3.1 years. In univariate analysis 4.5 % of the subjects receiving aspirin and 4.5 % of those not receiving aspirin died (HR 1.08, 0.85–1.39). Baseline characteristics were dissimilar in 25 of 31 of the covariates. In further analysis using matching by propensity score, 1,351 patients who were taking aspirin were at lower risk for death than 1,351 patients not using aspirin (4 % vs. 8 %, respectively; HR, 0.53; 95 % CI, 0.38–0.74; P = .002). After adjusting for the propensity for using aspirin, as well as other possible confounders and interactions, aspirin use remained associated with a lower risk for death (adjusted HR, 0.56; 95 % CI, 0.40–0.78; P < .001-Table 3.2). The patient characteristics associated with the most aspirin-related reductions in mortality were older age, known coronary artery disease, and impaired exercise capacity.

The major shortcoming with these aforementioned approaches is that one must know what the potential confounders are in order to match or adjust for them; and, it is the ***unknown confounders*** that represent a bigger problem. Another issue is that even if one suspects a confounder, one must be able to appropriately measure it. For example, socio-economic status (usually a combination of education and income) is a commonly addressed confounder; but, the definition of socio-economic status is an issue in which there is disagreement; and, which measures or cut-points to use is/are appropriate is frequently argued. The bottom line is that one can never perfectly measure all known confounders and certainly one cannot measure or perfectly match for unknown confounders. As mentioned, the strength of the RCT is that randomization (performed properly and with a large enough sample size) optimally balances both the known and unknown confounders between the interventional and control groups. But even with an RCT, randomization can be further compromised as will be discussed in some of the following chapters, and by the following example from "Student's" Collected Papers regarding the Lanarkshire Milk Experiment [20].

> *"Student" (i.e., the great William Sealy Gosset) criticized the experiment for it's loss of control over treatment assignment. As quoted: … Student's "contributions to statistics, in spite of a unity of purpose, ranged over a wide field from spurious correlation to Spearman's correlation coefficient. Always kindly and unassuming, he was capable of a generous rage, an instance of which is shown in his criticism of the* Lanarkshire *Milk Experiment. This was a nutritional experiment on a very large scale. For four months 5,000 school children received three-quarters of a pint of raw milk a day, 5,000 children the same quantity of pasteurized milk and 10,000 other children were selected as controls. The experiment, in Gosset's view, was inconclusive in determining whether pasteurized milk was superior in nutritional value to raw milk.*
>
> *This was due to failure to preserve the random selection of controls as originally planned. "In any particular school where there was any group to which these methods (i.e., of random selection) had given an undue proportion of well-fed or ill-nourished children, others were substituted to obtain a more level selection." The teachers were kind-hearted and tended to select ill-nourished as feeders and well-nourished as controls. Student thought that among 20,000 children some 200–300 pairs of twins would be available of which some 50 pairs would be identical-of the same sex and half the remainder nonidentical of the same sex. The 50 pairs of identicals would give more reliable results than the 20,000 dealt with in the experiment, and great expense would be saved. It may be wondered, however, whether Student's suggestion would have proved free from snags. Mothers can be as kind-hearted as teachers, and if one of a pair of identical twins seemed to his mother to be putting on weight…*

Missing Data

In 2008 the FDA requested that the National Research Council (NRC) convene an expert panel and to prepare a report that would be useful. The FDA that would address appropriate methods for analysis of missing data. Recall that the key feature of a RCT is the randomization process; and, this key feature is jeopardized when some of the outcome measures are missing. Missing data can seriously compromise the interpretations of clinical trials. A major source of missing data is the result of

Table 3.3a Eight ideas for limiting missing data in the design of clinical trials

Target a population that is not adequately served by current treatments and hence has an incentive to remain in the study
Include a run-in period (See discussion above regarding run-in periods)
Allow for a flexible treatment regimen
Shorten the follow up time so that participants are less likely to withdraw
Allow the use of rescue medications
For long term efficacy trials consider a withdrawal design
Consider an outcome that is not likely to lead to missing data
Consider add-on designs i.e. where a study treatment is added to existing therapies the patient may be receiving

Table 3.3b Eight ideas for limiting missing data in the conduct of clinical trials

Select investigators with good track records
Set acceptable rates for missing data and monitor during the course of the trial
Provide incentives to investigators and participants to continue the trial
Limit participant burden of data collection
Provide continued access to the trial medication after trial completion
Train investigators and study staff on the importance of trial continuation
Keep up to date contact information on trial participants
Assess the likelihood of participant continuation before enrollment

patients dropping out (discontinuing treatment) for any of a variety of reasons (adverse events, lost to follow up, lack of efficacy or tolerability etc). To the degree possible, these dropouts should be avoided, since there is no foolproof way to analyze data when there is significant (greater than 10 %?) missing data. Continuing to follow the patient after treatment discontinuation is one important step to reduce the degree of lost information. Little et al. summarized eight design ideas and eight ideas for the conduct of clinical trials for limiting missing data (Tables 3.3a and 3.3b [9]).

Since there is no universal method for handling missing data the best strategy is to avoid it. Statistical approaches to missing data will always involve unprovable assumptions, because there is always some uncertainty about the reasons that data is missing. The frequency of missing data is a result of patient dropouts the common reasons for which are: intolerability to the intervention, lack of intervention efficacy, or failure to attend designated appointments. Fleming has pointed out that there are only two reasons a patient can be off study; withdrawal of consent AND refusal to be followed or contacted, or the patient has achieved the required efficacy and safety end points [21]. He suggested six strategies to prevent missing data: first to distinguish nonadherence from nonretention; second to attempt to continue contact with the patient even if they have withdrawn from the study; third, adequately educate the patient during the informed consent process of the scientific relevance of the data they are providing; fourth, protocols should not give a false sense of being able to correct for missing data with statistical approaches; fifth, protocols should specify targeted levels of data capture; and, sixth, forms and procedures for data collection should be formulated to reduce the likelihood of missing data. The use of run-in periods is an additional strategy, along with the use of flexible doses may be helpful as well.

Losses to follow-up (see below) using the last observation carried forward or baseline observation carried forward analysis is likely to overestimate and/or bias the outcome (since patients lost to follow-up more frequently are not benefitting from the intervention). Imputing the worst possible outcome might underestimate the benefit of the intervention. Missing data can be viewed in several ways. The ideal is if the missing data is "missing completely at random" (MCAR). This is an assumption that is unlikely to hold in most clinical trials because it presumes that the missing data are unrelated to the study variables (an unlikely scenario). A more realistic condition is missing at random (MAR, this might be better stated as missing "mostly" at random), or missing not at random (MNAR).

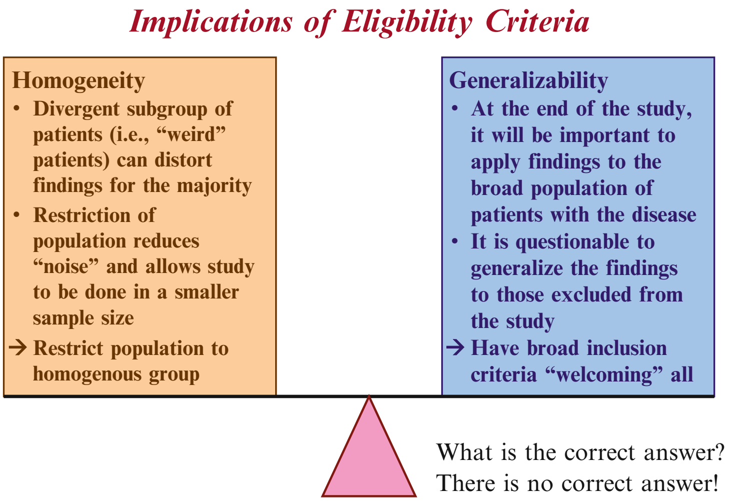
Because some missing data that does occur in almost every study, and each clinical trial has its own set of challenges, the NRC panel did list four general approaches: complete-case analysis, single imputation methods, estimating-equation methods and methods based on a statistical model. There is no single correct method for handling missing data, as all methods require that untestable assumptions be made. Discussion of these are beyond the scope of this book, but briefly, complete-case analysis simply excludes participants with missing data while with imputation, a single value is filled in for each missing value by using such methods as last observation carried forward or the baseline value carried forward. With estimating-equation methods, cases are weighted based upon the estimate of probability of an outcome being observed. As to the statistical modeling, approaches such as prior probabilities (Bayesian Methods) and multiple imputation where multiple sets of plausible values for missing data are used. Missing data can occur, of course, at random, or there can be differential loss of data, a more important consideration when missing data is assessed. Little et al. outlined six principles for drawing inferences from incomplete data [9].

1. Consider if the missing values are meaningful for analysis
2. Consider a possible causal pathway and how missing data might influence it
3. Consider why data are missing
4. Decide on a set of assumptions about the mechanism for missing data
5. Conduct a statistically valid analysis based on the above
6. Conduct a sensitivity analysis, a statistical technique that attempts to determine how changes in one variable will impact the target variable [7].

Complications of Eligibility Criteria

All generalizations are false, including this one (Mark Twain)

In every study there are substantial gains in statistical power by focusing the intervention in a homogenous patient population likely to respond to treatment, and to exclude patients that could introduce 'noise' by their inconsistent responses to treatment.



Implications of Eligibility Criteria

Homogeneity

- **Divergent subgroup of patients (i.e., "weird" patients) can distort findings for the majority**
- **Restriction of population reduces "noise" and allows study to be done in a smaller sample size**

→ Restrict population to homogenous group

Generalizability

- **At the end of the study, it will be important to apply findings to the broad population of patients with the disease**
- **It is questionable to generalize the findings to those excluded from the study**

→ Have broad inclusion criteria "welcoming" all

What is the correct answer?
There is no correct answer!

Fig. 3.2 The balance of conflicting issues involved with patient selection

Conversely, at the end of a trial there is a need to generalize the findings to a broad spectrum of patients who could potentially benefit from a superior treatment. These conflicting demands introduce the issue of balancing the inclusion/exclusion (eligibility criteria) such that the enrolled patients are as much alike as possible; but, on the other hand to be as diverse as possible in order to be able to apply the results to the more general population (i.e. generalizability). Figure 3.2 outlines this balance. What is the correct way of achieving this balance? There really is no correct answer, there is always a tradeoff between homogeneity and generalizability; and each study has to address this, given the availability of subjects, along with other considerations. This process of sampling represents one of the reasons that scientific inquiry requires reproducibility of results, that is, one study generally cannot be relied upon to portray 'truth' even if it is a RCT. The process of sampling embraces the concept of generalizability. The issue of generalizability is nicely portrayed in a video entitled 'A Village of 100' [22]. If one wanted to have a representative sample of the world for a study, this video (although predominately focused upon tolerance and understanding), is an excellent way of understanding the issue of generalizability. The central theme of the video asks the question 'if we shrunk the earth's population to a village of precisely 100 people, with all existing ratios remaining the same, what would it look like?' To paraphrase, if we maintained the existing ratios of the earth's population in a study of 100 people, what would our sample look like? The answer – there would be 57 Asians, 21 Europeans, 14 from the Western Hemisphere, 51 females and 49 males, 70 non-white and 30 white, 70 non-Christians and 30 Christians, 89 heterosexuals, 50 % of the worlds wealth would belong to 6 citizens of the USA, 80 would live in sub-standard housing, 70 would be unable to read (a potential problem with IRB approval), 50 would be malnourished, one would have a college education, and 4 would own a computer. When is the last time a study had a population representative of the Village of 100?

Table 3.4 Birmingham v Framingham: comparison of key variables

	Birmingham	Framingham
Population	242,800	62,910
% African American	73.5	5.1
Age		
25–44	30	35
45–64	20	33
65–>	14	13
Median income $	26,700	55,300
Education %		
<High school	25	13
High school	28	23
>High school	48	64
CVD rate	528–582	336–451

For an example of sampling issues, most of the major studies assessing the efficacy of the treatment of extracranial atherosclerosis with endarterectomy had excluded octogenarians on the basis that this patient population may have a response to the challenges of surgery that is different than their younger counterparts [23, 24]. Exclusion of these patients may have contributed to the successful completion of 'positive' trials (finding a benefit for the then new treatment – endarterectomy). However, now that the trials are complete, there is not 'level 5' evidence (data that is a result from RCTs) to guide the management of octogenarians with extracranial atherosclerosis, one of the subpopulations where the need for this information is important. In the absence of this information, thousands of endarterectomies are performed in this older patient population each year under the assumption that the findings from a younger cohort are generalizable to those at older ages. For another example, let's presume that in a multicenter trial that included Framingham Massachusetts, and Birmingham, Alabama, that a representative sample of each was recruited into a study. The makeup of the sample from each is illustrated in Table 3.4. As one can see, there are significant differences in the representative sample populations, and these differences could affect not only the success of the intervention but could also confound its relationship.

Efficacy vs. Effectiveness

Another limitation of RCTs is that they are designed to test safety and efficacy (i.e. does the drug work under optimal circumstances?) and not to answer questions about the effectiveness of a drug, the more relevant question for clinicians and economic analysts (i.e. does the drug work under ordinary circumstances of use?). Thus, the increased use of effectiveness trials has been suggested, to more closely reflect routine clinical practice. Effectiveness trials use a more flexible dosage regimen, and generally a 'usual care' comparator instead of a placebo comparator. Two approaches to this more 'real world trial' is the phase 4 trial (see Chap. 5) or

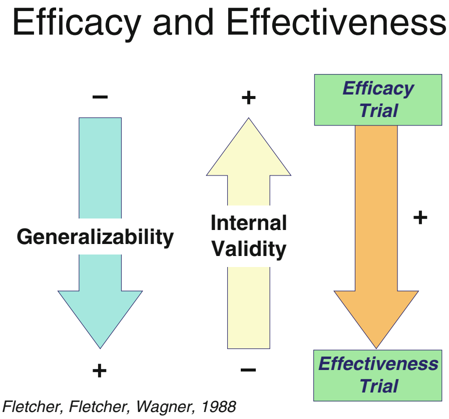


Fig. 3.3 The "Trade-off" between efficacy vs. effectiveness

the prospective, randomized, open-label, blinded end-point – PROBE-Trial. The PROBE Trial is further discussed in the next section entitled "Degree of Masking"). As to phase 4 trials, they are surrounded by some controversy as well. Figure 3.3 compares efficacy and effectiveness trials in terms of some of their more important variables.

Patient Compliance

Run-in Periods

Another issue surrounding RCTs, and one that is almost unique to clinical trials, is the use of run-in periods and their impact on who is eligible to be randomized. Pre-randomization run-in periods are frequently used to select or exclude patients in clinical trials, but the impact of run-in periods on clinical trial interpretation and generalization has not been systematically studied. The controversy regarding run-in periods also addresses the issue of efficacy vs. effectiveness, as the run-in period allows one to exclude patients that are potentially less compliant, or do not tolerate placebo (or whatever other intervention is used in an active comparison group). Although this issue has not been systematically studied, intuitively one can see that the potential for over-estimating the impact of an investigational drug is present when run-in periods are utilized, as the run-in period will likely exclude patients from the study who would not have ideally responded.

A study can achieve high compliance in at least three general ways: designing a simple protocol (complexity makes compliance more difficult); the use of compliance aids such as automatic reminders, telephone calls, calendars, etc; or by selecting subjects based upon pre-study or pre-randomization compliance. Of course, high compliance is a desirable characteristic of any research. High compliance attenuates the argument of whether to use intention to treat vs. compliance only as the primary analysis. Also, high compliance will optimize the studies power as the "diluting" effect of non-compliers will not be manifest (all other things being equal). While the run-in period increases the proportion of compliers in the trial, it may introduce important differences in the outcomes, particularly if compliers and non-compliers are inherently different in the way they would respond to the intervention of interest. Thus, the effect of run-in periods on generalizability should be considered carefully before implementation. Lang [25] has listed some recommendations for helping to decide whether to use a run-in as part of a clinical trial, including:

1. consider a run-in whenever the contact between study staff and participants is low
2. consider a run-in period for a primary prevention trial because compliance is likely to be more difficult compared to therapeutic trials
3. For any trial, list the key features of the study protocol and see which features compliance could be directly tested prior to randomization
4. before using active agents during a run-in, consider both the expected frequency of occurrence of side effects and the postulated effect of the agent on the outcome of interest
5. all trials can use any available pre-randomization period for the simultaneous purpose of characterizing patients and evaluating compliance, whether of not the compliance information will be used for exclusions

In fairness, as Franciosa points out, clinicians use variants of run-in periods to treat their patients, such as dose titration, or challenge dosing (e.g. using small doses of ACE Inhibitors to rule out excessive responders) [26]. Pablos-Mendez et al. analyzed illustrative examples of reports of clinical trials in which run-in periods were used to exclude non-compliant patients, placebo responders, or patients that could not tolerate or did not respond to active drug [27].

Thus, the use of run-in periods is another reason that the results of RCTs may not accurately portray what the drugs overall effectiveness will be. What can be said is that there does need to be more focus on the details of run-in periods, and as is true of most things the researcher does in designing and implementing a clinical trial, judgments have to be made regarding the best approach to use regarding inclusions and exclusions, as well as judging what the impact of the run-in period is on the ultimate interpretation of a clinical trial. Ultimately, from the perspective of internal validity, it is better to exclude participants before randomization than have participants lost to follow up, cross between study groups, or become non-adherent to intervention protocols after randomization.

Recruitment and Retention

Nothing is more critical to the success of a clinical trial than the recruitment and retention of subjects. As will be discussed in more detail in Chap. 8, there are a number of reasons for failure of the recruitment process including: delayed start-up, and inadequate planning, In terms of patient/subject retention, there are arguably differences in the handling of clinical patients in contrast to research subjects (although this could and perhaps should be challenged). Losses-to-follow-up need to be kept to a minimum and is discussed later in this chapter.

Degree of Masking (Blinding)

Although the basic concept of clinical trials is to be at equipoise, this does not change the often pre-conceived 'suspicion' that there is a differential benefit of the investigational therapy (e.g. the investigational drug is better than placebo). Thus, if study personnel know the treatment assignment, there may be differential vigilance where the supposed 'inferior group' is more intensively monitored (e.g. 'are you certain you have not had a problem?' they might ask). In this case, unequal evaluations can provide unequal opportunities to differentially 'discover' events. This is why the concept of double-blinding (masking) is an important component of RCTs. There is an argument about which term-blinding or masking-is most appropriate [28], and Fig. 3.4 portray's a humorous example of this argument. But, one cannot always have a double-blind trial, and some would argue that double-blinding distances the trial from a 'real-world' approach. An example where blinding is difficult to achieve might be a surgical vs. medical intervention study where post operative patients may require additional follow-up visits, and each visit imparts an additional opportunity to elicit events. That is, it has been said that 'the patient

Fig. 3.4 A humorous example of blinding (masking) (With permission from Schulz and Grimes [28])

cannot have a fever if the temperature is not taken,' [29] and for RCTs, events cannot be detected without patient contact to assess outcomes.

Of course, masking is not always possible and examples include studies that: might evaluate residual surgical wounds, studies involved with cycling hormone replacement, studies requiring serum (or other) assay or physical measurement, studies that involve participant participation in the treatment (i.e., low fat diet, exercise, etc). Common approaches to these examples are to at least mask the rater (adjudicator), or to move toward a totally objective outcome (e.g. death), or to use an independent observer who does not know treatment to assess outcome. In an effort to study the impact of adjudicator blinding on outcomes, Parmar et al. assessed the effect of blinding race and geography on outcomes ascertainment in an observational study [28]. The primary characteristics of interest were race and geography, and the prespecified acceptable agreement rate between adjudicators was set at >80 %. They selected 116 suspected cardiovascular events that underwent adjudication with usual blinding. At least 3 months later, cases were readjudicated without blinding race and geographic location of the patient, and differences in outcomes ascertainment was assessed using Cohen's κ statistic and agreement rates. Agreement between the blinded and unblinded reviews was good to excellent for all four outcomes. κ statistics were 0.80 (chest pain), 0.85 (heart failure), 0.86 (revascularization) and 0.74 (MI) ($p<0.0001$ for all). Within each outcome, agreement rates were similar for race and geographic groups (agreement 83–100 %). The authors concluded that in observational studies, blinding medical record review for outcomes ascertainment for some types of patient characteristics may be an unwarranted expense.

In order to realize a more 'real-world' approach to clinical trials, the prospective randomized open-label blinded endpoint design (PROBE design) was developed. Randomization is used so that this important component of study design is retained. By using open-label therapy, the drug intervention and its comparator can be clinically titrated as would occur in a doctor's office. Of course, blinding is lost here, but only as to the therapy. In a PROBE design, blinding is maintained as to the ascertainment of the outcome. To test whether the use of open-label vs. double-blind therapy affected outcomes differentially, a meta analysis of PROBE trials and double-blind trials in hypertension was reported by Smith et al. [30]. They found that changes in mean ambulatory blood pressure from double-blind controlled studies and PROBE trials were statistically equivalent.

Selection of Comparison Groups

As the story goes a clinical researcher meets someone on the street who asks "how do you do?" The researcher answers "compared to whom?" When addressing the validity of an outcome difference compared to some control group, it is crucial that the control group be clearly defined. Sometimes studies assess a new (investigational) treatment versus an approved (standard) active treatment (i.e. to assess if the old 'standard' treatment should be replaced with the new treatment), in other cases, studies are assessing if a new treatment should be added (not replacing, but rather supplementing), current treatment. In this latter case, the comparison of interest is the outcome of patients with and without the new treatment. In this instance, masking can only be accomplished by the use of a double-blind technique. Traditionally, placebo treatment has been used as the comparator to investigational treatments, and has been one of the standards of clinical trials.

The use of the placebo comparator has more and more been the subject of ethical concerns. In addition to ethical issues involved with the use of placebos, there are other considerations raised by the use of placebo-controls. For example, an important lesson was learned from the Multiple Risk Factor Intervention Trial (MRFIT) regarding the use and analysis of the placebo control group, which might best be summed up with the question 'why it is important to watch the placebo group?' [31]. MRFIT screened 361,662 patients to randomize high-risk participants (using the Framingham criteria existent at that time) to special intervention (n = 6428) and usual care (n = 6438) with coronary heart disease mortality as the endpoint. The design of this well-conducted study assumed that the risk factor profile of those receiving 'special treatment interventions' would improve, while those patients in the 'usual care' group would continue their current treatments and remain largely unaffected. The special intervention approaches in MRFIT were quite successful, and all risk factor levels were reduced. However, there were also substantial and significant reductions observed in the control group. That both treatment and control groups experienced substantial improvements in their risk factor profile translated to almost identical CHD deaths during the course of the study. Why did the control group fare so well? Several phenomena may have contributed to the improvement in the placebo-control group. First, is the Hawthorne effect, which suggests that just participating in a study is associated with increased health awareness and changes in risk factor profile, irrespective of any intervention [32]. In addition, for the longer-term trials, there are changes in the general population that might alter events. For example, randomization in MRFIT was conducted during the 1980s, a period when health awareness was becoming more widely accepted in the USA, and likely beneficially affected the control group.

Although the ethics of placebo controls is under scrutiny, another principal regarding the placebo-control group is that sometimes being in the placebo group isn't all that bad. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study was launched in 1994 [33]. By the early 1990s there was mounting clinical epidemiologic evidence of reduced cancer risk associated with a higher intake of antioxidants. Treatment with vitamin E and beta carotene were considered unlikely to be harmful, and likely to be helpful; and, the question was asked whether antioxidants could reduce lung cancer-even in smokers. A double-blind, placebo-controlled RCT was launched with a 2 x 2 factorial design (see Chap. 4), and over 7,000 patients in each cell. No benefit was seen with either therapy, but compared to placebo; a disturbing worsening trend was observed in the beta-carotene treated compared with the placebo group.

Table 3.5 Different definitions of “normal”

Property	Term	Consequence of application
Distribution shape	Gaussian	Minus values
Lie within preset %	Percentile	Normal until workup
No additional risk	Risk factor	Assumes altering risk factor improves risk
Societal or political	Culturally desirable	Raises the role of society in medicine
A range before test suggests no disease	Diagnostic	Need to know the predictive value in ones own practice
Therapy beneficial	Therapeutic	New therapies alter this

Frequently, the comparison group or control group is a so called ‘normal’ population. Inherent to this concept is ‘what is normal?’. A wit once opined that ‘a normal person is one who is insufficiently tested’. Interestingly, there are a number of scientific definitions of normal (see Table 3.5). One definition of normal might be someone who fits into 97 % of a Gaussian Distribution, another that normal lies within a preset percentile of a laboratory value or values. Other definitions exist, suffice it to say, whatever definition is used it needs to be clearly identified.

Analytic Approach

Intention to Treat and Per-Protocol Analysis

There are three general analytic approaches to clinical trials; intention-to-treat (ITT) analysis (or analysis as randomized), compliers only (or per-protocol) analysis, and analysis by treatment received. Probably the least intuitive and the one that causes most students a problem is ITT. ITT was derived from a principle called the pragmatic attitude [34]. The concept was that one was to compare the effectiveness of the *intention* to administer treatment A vs. the *intention* to administer treatment B, i.e. the comparison of two treatment policies rather than a comparison of two specific treatments. With ITT, everyone assigned to an intervention or control arm is counted in their respective assigned group, whether they ultimately receive none of the treatment, or somewhat less than the trial directed. For example, if in a 1-year trial, a patient is randomized to receive an intervention, but before the intervention is administered, they drop out (for whatever reason) they are analyzed as if they received the treatment for the entire year. The same applies if the patient drops out at any time during the course of the study. Likewise, if it is determined that the patient is not fully compliant with treatment, they are still counted as if they were. In fact, whether there is compliance, administrative, or protocol deviation, patients once randomized are counted as if they completed the trial. Most students initially feel that this is counter-intuitive. Rather the argument would be that one is really interested in what would happen if a patient is randomized to a treatment arm and they take that treatment for the full trial duration and are fully compliant – this, one would argue, gives one the real information needed about the optimal effect of an intervention (this, by the way, is a description of the compliers only analysis). So why is ITT the scientifically accepted primary analysis for most clinical trials? As mentioned before, randomization is arguably one of the most important aspects of clinical trial design. If patients once randomized to a treatment are not included in the analysis, the process of randomization is compromised. It is not a leap of faith to wonder if patients dropping out of the intervention arm might be different than the patients dropping out of a control arm. Thus, if ITT is not used, one loses the assurance of equal distribution of unknown confounders between the treatment groups, and this thereby tarnishes the basis of randomization. One example of the loss of randomization if ITT is not used might be differential dropouts between the intervention and control arm for adverse events. Also, if patients with more severe disease are more likely to dropout from the placebo arm; or conversely patients who are older, dropout more frequently from the placebo arm thereby removing them from the analysis, this could result in an imbalance between the two comparison groups. Another argument for ITT is that it provides for the most conservative estimate of the intervention effect (if the analysis includes patients that did not get the entire treatment regimen and the regimen is beneficial, clearly the treatment effect will be diluted). Thus, if using ITT analysis reveals a benefit, it adds to the credibility of the effect measure. Of course, one could argue that one could miss a potentially beneficial effect if the intervention effect is diluted. In summary, ITT protects against bias, protects the statistical integrity of the trial, and protects the randomization process.

In the compliers only analysis, the patients that complete the trial and comply fully with that treatment are analyzed. The problem is that if a beneficial effect is seen, one can wonder what the loss of randomization (and thereby equality of confounders between groups) means to that outcome, particularly if an ITT analysis does not demonstrate a difference. The loss of randomization and the loss of balanced confounders between the treatment and control groups is exemplified by an analysis of the Coronary Drug Project, where it was determined that poor compliers to placebo had a worse outcome than good compliers to placebo [35]. This would suggest that there are inherent differences in patients who comply vs. those who do not, and this could differentially be the cause of dropout. The Coronary Drug Project was a trial aimed at comparing clofibrate with placebo in patients with previous myocardial infarction with the outcome of interest being mortality. Initially reported as a favorable intervention (there was a 15 % 5 year mortality in the clofibrate compliers only analysis group, compared to a 19.4 % mortality in the placebo group-$p < .01$); while with ITT analysis there was essentially no difference in outcome (18.2 vs. 19. 4 %–$p < .25$). Given the differences in outcome between placebo compliers and placebo non-compliers, one can only assume the same for the investigational drug group. Likewise, the Anturane Reinfarction Trial was designed to compare anturane with placebo in patients with a prior MI and in whom mortality was the outcome of interest [36]. One thousand six hundred and twenty nine patients were randomized 817 to placebo and 812 to anturane (71 patients were later excluded because it was determined that they did not meet eligibility criteria).

The study initially reported anturane as a favorable intervention (although the $p < .07$), but when the 71 ineligible randomized patients were included in the analysis the $p = 0.20$. Again further analysis demonstrated that in the anturane ineligible patients, overall mortality was 26 % compared to the mortality in the anturane eligible patients that was 9 %.

If one considers the common reasons for subjects not being included in a study, ineligibility is certainly one. In addition, subjects may be dropped from a trial for poor compliance, and/or adverse drug events; and, patients may be excluded from analysis due to protocol deviations or being lost to follow up. Some of the reasons for ineligibility are protocol misinterpretations, clerical error, or wrong diagnosis at the time of randomization. Sometimes the determination of ineligibility is above question (e.g. the patient fell outside of the studies predetermined age limit) but frequently ineligibility requires judgment. The Multicenter Investigation of the Limitation of infarct Size (MILIS) study is an example of this latter concept. MILIS compared propranolol, hyaluronidase, and placebo in patients with early acute MI, in order to observe effects on mortality. Subsequently, some patients were deemed ineligible because the early diagnosis of MI was not substantiated. But, what if the active therapy actually had an effect on preventing or ameliorating the MI? The problem with not including patients in this instance is that more patients could have been withdrawn from the placebo group compared to the active therapy group, and as a result, interpretation of the data would be altered.

Of course, as is true of most things in clinical research there is not just one answer, indeed, one has to carefully assess the trial specifics. For example, Sackett and Gent cite a study comparing heparin to streptokinase in the treatment of acute myocardial infarction [37]. The ITT analysis showed that streptokinase reduced the risk of in-hospital death by 31 % ($p = 0.01$). However, eight patients randomized to the heparin group died after randomization, but before they received the heparin. Analysis restricted to only those who received study drug decreased the benefit of streptokinase (and increased the p value).

In summary, ITT is the most accepted (e.g. by most scientists and the FDA) as the analysis of choice for clinical trials. This is because ITT assures statistical balance (as long as randomization was properly performed), it ‘forces’ disclosure of all patients randomized in a trial, and most of the arguments against ITT can be rationally addressed.

Analysis-As-Treated is another analytic approach that addresses not the group to which the patient was randomized and not compliers only, but what the patient actually received. This analytic approach is utilized most often when patients cross over from one treatment arm to the other; and, this occurs most often in surgical vs. medical treatment comparisons. For example, patient’s randomized to medical treatment (vs. coronary artery bypass surgery) might, at some time during the study, be deemed to need the surgery, and are thus crossed over to the surgical arm and are then assessed as to the treatment they received (i.e. surgery). Like compliers only analysis, this might be an interesting secondary analytic technique, but shares many of the same criticisms discussed earlier for compliers-only analysis. In addition, because such trials cannot easily be double-blinded, even greater criticism can be leveled against this analytic approach compared to compliers-only analysis. In addition, statistical testing with this analysis by treatment received, is more complicated, not only by the crossovers, but by the inherent nature of the comparison groups. In comparison trials of 1 drug vs. placebo, for example, it is reasonable to assume that if the drug is superior to placebo (or an active control) patients in the drug group will average fewer events in the follow-up period. When this is displayed as survival curves, the survival curves will increasingly separate. In trials comparing surgical to medical therapy, the aforementioned approach may not be reasonable. For example, if patients randomized to surgery have a high early risk (compared to the non-surgical group) and a lower risk later, these risks may cancel and be similar to the number of events under the null hypothesis of no difference between groups. The issue of comparing surgical and non-surgical therapies in clinical trials has been nicely summarized by Howard et al. [38].

Subgroup Analysis

As pointed out by Assmann et al., most clinical trials collect substantial baseline information on each patient in the study [39]. The collection of baseline data has at least four main purposes: (1) to characterize the patients included in the trial, i.e. to determine how successful randomization was (2) to allow assessment of how well the different treatment groups are balanced, (3) to allow for analysis per treatment, (4) to allow for subgroup analysis in order to assess whether treatment differences depend on certain patient characteristics. It is this 4th purpose that is perhaps the most controversial because it can lead to 'data dredging' or has some wits have opined, 'if you interrogate the data enough, you can get it to admit to anything'. For example, Sleight and colleagues, in order to demonstrate the limitations of subgroup analysis, performed subgroup analysis in the ISIS-2 trial by analyzing treatment responses according to the astrological birth sign of the subject [40]. This analysis suggested that the treatment was quite effective and statistically significant for all patients except those born under the sign of Gemini or Libra. The validity of any subgroup observation tends to be inversely proportional to the number of subgroups analyzed. For example, for testing at the 5 % significance level ($p \leq .05$) an erroneous statistically significant difference will be reported (on average) 5 % of the time (i.e. false + rate of 5 %). But, if 20 subgroups are analyzed, the false positive rate would approach 64 % (Table 3.6).

It is true, that meaningful information from subgroup analysis is restricted by multiplicity of testing and low statistical power and that surveys on the adequacy of the reporting of clinical trials consistently find the reporting of subgroup analyses to be wanting. Most studies enroll just enough participants to ensure that the primary efficacy hypothesis can be adequately tested, and this limits the statistical ability to find a difference in subgroup analyses; and, the numbers of subjects available for subgroup analysis is further compounded by loss of compliance, the need for adjustments for multiple testing, etc. Some have taken this to mean that subgroup

Table 3.6 Approximate number of False Positives (FP) occurring with multiple subgroup analyses

No. of tests	Probability of 1 FP	Probability of 2 FPs	Probability of 3 FPs
1	0.05	0.01	0
2	0.10	0.02	0
3	0.14	0.025	0
5	0.23	0.03	0
10	0.40	0.05	0.01
20	0.64	0.10	0.10

analyses are useless. When results from a subgroups analysis are at variance from the overall group outcome, the results are still likely to be true if the subgroup is large, they are pre-specified rather than *post hoc* (i.e. 'after the fact') and they are of limited number (not all post hoc analyses are subgroup analyses, but arguably most are). At the least, whether pre-specified or *post hoc,* subgroup analyses serve to generate questions for subsequent trials, and should not be interpreted as "truth". An exception to this latter principal, is when it comes to safety, here subgroup analyses might "carry more weight". An example of a post-hoc analysis that was "accepted" is the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) study where LIPITOR 80 mg vs. placebo was administered in 4,731 subjects without CHD who had a stroke or TIA within the preceding 6 months [41]. A higher incidence of hemorrhagic stroke was seen in subgroup analysis in the LIPITOR 80 mg group compared to placebo. Subjects with hemorrhagic stroke on study entry appeared to be at increased risk for hemorrhagic stroke. As a result, Pfizer revised the US Prescribing Information for atorvastatin to include a precaution for its use of 80 mg in patients with a prior history of stroke.

What can be said is that if subgroup analysis is used and interpreted carefully, it can be useful. Even among experts, opinions range from only accepting pre-specified subgroup analyses supported by a very strong *a priori* biological rationale, to a more liberal view in which subgroup analyses, if properly carried out and interpreted, are permitted to play a role in assisting doctors and their patients to choose between treatment options. In reviewing a report that includes subgroup analyses, Cook et al. suggest addressing the following issues (Table 3.7): (1) were the subgroups appropriately defined, (that is, be careful about subgroups that are based upon characteristics measured after randomization e.g. adverse drug events may be more common as reasons for withdrawal from the active treatment arm whereas lack of efficacy may be more common in the placebo arm); (2) were the subgroup analyses planned before the implementation of the study (in contrast to after the study completion or during the conduct of the study); (3) does the study report include enough information to assess the validity of the analysis e.g. the number of subgroup analyses; (4) do the statistical analyses use multiplicity and interaction testing; (5) were the results of subgroup analyses interpreted with caution; (6) is there replication of the subgroup analysis in another independent study; (7) was a dose-response relationship demonstrated in the subgroup; (8) was there reproducibility of the observation within individual sites; and (9) is there a biological explanation.

Table 3.7 Considerations regarding subgroup analyses

Was there potential for patient misclassification
Was the analysis approach Intention-To-Treat
Were subgroups planned *a priori*
Was the subgroup analysis based on trial or biological data
Was there adequate power for subgroup analysis
What are the total number of subgroups analyzed
Are there adjustments for multiple testing
Are there tests for interaction
Are subgroup results emphasized above primary analyses
Are the subgroup analyses placed in proper biological and prior trial data perspective
Are *a priori* analyses distinguished from *a posteriori* analyses

Table 3.8 Goal of RCTs and their relation to hypothesis testing

RCT goal	Superiority	Equivalence
Null hypothesis	New = Old	New < Old + δ
Alternative hypothesis	New $\overset{\diagdown}{=}$ Old	New = Old + δ

δ is the margin in which the point estimate falls

Traditional Versus Equivalence Testing (Table 3.8)

Most clinical trials have been designed to assess if there is a difference in the efficacy to two (or more) alternative treatment approaches (with placebo usually being the comparator treatment). There are reasons why placebo-controls are preferable to active controls, not the least of which is the ability to distinguish an effective treatment from a less effective treatment. However, if a new treatment is considered to be equally effective but perhaps less expensive and/or invasive, or a placebo-control is considered unethical, then the new treatment needs to be compared to an established therapy and the new treatment would be considered preferable to the established therapy, even if it is just as good (not necessarily better) as the old (Table 3.9). The ethical issues surrounding the use of a placebo-control and the need to show a new treatment to only be as 'good as' (rather than better) has given rise to the recent interest in equivalence or non-inferiority testing. With traditional (superiority) hypothesis testing, the null hypothesis states that 'there is no difference between treatment groups (i.e. New = Old or placebo or standard therapy). Rejecting the null, then allows one to definitively state if one treatment is better (or worse) than another (i.e. New > or < Old). The disadvantage is if at the conclusion of an RCT there is not evidence of a difference, one cannot state that the treatments are the same, or as good as one to the other, only that the data are insufficient to show a difference. That is, when the null hypothesis is not accepted, it is simply the case where it cannot be rejected. The appropriate statement when the null hypothesis is not rejected (accepted) is 'there is not sufficient evidence in these data to establish if a difference exists.'

Table 3.9 Reasons for choosing noninferiority over superiority designs

Comparing new treatment with active control instead of placebo	Unethical to use placebo group in controlled study when there's an established treatment
New treatment not better in primary end point; better in secondary end points	Although no difference between primary efficacy outcomes, difference in secondary end points such as adverse events, quality of life
New treatment not better in primary end point; overall efficiency is better	Non-inferiority in effectiveness and safety; clear superiority in incurred cost produces and overall efficiency
The new treatment can be non-inferior and superior	Non-inferiority testing can be complemented by superiority testing in one study without need for adjustments

Equivalence testing in essence 'flips' the traditional null and alternative hypotheses. Using this approach, the null hypothesis is that the new treatment is worse than the old treatment (i.e. New < Old); that is, rather than assuming that there is no difference, the null hypothesis is that a difference exists and the new treatment is inferior. Just as in traditional testing, the two results available from the statistical test are (1) reject the null hypothesis, or (2) failure to reject the null hypothesis. However, with equivalence/noninferiority testing rejecting the null hypothesis is making the statement that the new treatment is not worse than old treatment, implying the alternative, that is 'that the new treatment is **as good** as the old' (i.e. New = Old). Hence, this approach allows a definitive conclusion that the new treatment is as good as the old.

One caveat is the definition of 'as good as,' which is defined as being in the 'neighborhood' or having a difference that is so small that it is to be considered clinically unimportant (generally, effects within ±2 % – this is known as the equivalence or noninferiority margin usually indicted by the symbol δ). The need for this 'neighborhood' that is considered 'as good as' exposes the first shortcoming of equivalence testing – having to make a statement that 'I reject the null hypothesis that the new treatment is worse than the old, and accept the alternative hypothesis that it is as good – *and by that I mean that it is within at least 2 % of the old*' (the wording in italics are rarely included in the conclusions of a manuscript). A second disadvantage of equivalence/noninferiority testing is that no definitive statement can be made that there is evidence that the new treatment is better or worse. Just as in traditional testing, one never accepts the null hypothesis – one only fails to reject it. Hence if the null is not rejected, all one can really say is that there is ***insufficient evidence in these data*** that the new treatment is as good as the old treatment. Another problem with equivalence/noninferiority testing is that one has to rely on the effectiveness of the active control obtained in previous trials, and on the assumption that the active control would be equally effective under the conditions of the present trial.

An example of an equivalence trial is the Controlled ONset Verapamil INvestigation of Cardiovascular Endpoints study (CONVINCE), a trial that also raised some ethical issues that are different from those usually involved in RCT's [42]. CONVINCE was a large double-blind clinical trial intended to assess the equivalence of verapamil and standard therapy in preventing cardiovascular disease-related events in hypertensive patients. The results of the study indicated that the verapamil preparation was not equivalent to standard therapy because the upper bound of the 95 % confidence limit (1.18) slightly exceeded the pre-specified boundary of 1.16 for equivalence. However, the study was stopped prematurely for commercial reasons. This not only hobbled the findings in terms of inadequate power, it also meant that participants who had been in the trial for years were subjected to a 'breach in contract'. That is, they had subjected themselves to the risk of an RCT with no ultimate benefit. There was a good deal of criticism borne by the pharmaceutical company involved in the decision to discontinue the study early. Parenthetically, the company involved no longer exists.

In the past, some separated equivalence testing and non-inferiority testing. The question posed by non-inferiority testing being slightly different in that one is asking whether the new intervention is simply not inferior to the comparator (i.e. New $\not<$ Old). One potential advantage of this approach is that statistical significance could be only 'one-tailed' since there is no implication that the analysis is addressing whether the new treatment is better or as good as, only that it is not inferior. There is a good deal of disagreement regarding this latter issue, so that most use the two (equivalence and noninferiority) approaches interchangeably. Weir et al. utilized the non-inferiority approach in evaluating a comparison of valsartin/hydrochlorthiazide (VAL/HCTZ) with amlodipine in the reduction of mean 24-h diastolic BP (DBP) [43]. Noninferiority of the VAL/HCTZ combination to amlodipine was demonstrated, and fewer adverse events were noted with the combination treatment as well. The null hypothesis for this analysis was that the reduction in mean 24-h DBP from baseline to the end of the study with VAL/HCTZ was ≥3 mmHg less (the non-inferiority margin) compared with amlodipine. Again, a caveat has been recently raised by LeHenanff et al. and Kaul et al. [44, 45]. LeHananff et al. [45] reviewed studies published between 2003 and 2004 that were listed as equivalence or noninferiority, and noted a number of deficiencies, key among them being the absence of a stated equivalence or non inferiority margin [45].

Equivalence/non-inferiority trials are further discussed in Chap. 4.

Losses to Follow Up (See also Discussion of Missing Data, Above)

Patients who are lost-to-follow-up are critical in clinical trials and are particularly problematic in long-term trials. Patients lost to follow-up might be regarded as having had poor results (that is assumed that they experienced treatment failure); so if there are sufficient numbers of them, trial results can be skewed to less of an effect, even if, in truth, they did not have poor results. If, in the different study arms, there are equal numbers lost to follow-up, and they are lost for the same reasons, lost to follow up would not be as critical, but this is unlikely to occur. Section 4.3.4 of the ICH E-6 Good Clinical Practice: Consolidated Guidance reads, "*Although a subject is not obliged to give his/her reason(s) for withdrawing prematurely from a trial, the investigator should make a reasonable effort to ascertain the reason(s),* *while fully respecting the subject's rights*" This excerpt expresses the need for physicians associated with clinical research trials to make a first-hand effort to contact patients who are lost-to-follow-up. In doing so pharmaceutical companies not only look out for the best interest of the patients who enroll in their clinical research trials, but also protect the data outcome of their clinical trials.

Of course, in ITT analysis, patient's lost-to-follow-up is still counted, but the argument is how to count them. Some would argue that it is appropriate to count them as poor outcomes since this will give the most conservative result, while others argue that since their outcome is not known, they should not be counted. In fact, there is little data reported on the actual impact on a study result of patients lost to follow up. In one study, Joshi et al. did address this issue in a long-term follow-up (up to 16 years of follow-up) of patients who had undergone knee arthroplasty. With the concerted effort of full-time personnel and a private detective, all 123 patients initially lost to follow-up were traced. Patients cited a variety of reasons why they did not attend follow-up visits, including: change of residence, inability to travel, displeasure with the physician or staff, financial constraints, satisfaction with the results so that they did not feel follow-up was necessary and poor results. They also found that more women than men were lost to follow-up. A few companies have developed methods of locating and contacting patients that are lost-to-follow-up and processes of handling patient information. These are options that pharmaceutical companies can use to find patients that have become lost-to-follow-up. These lost to follow-up patient locate systems use customized programmed software systems, as well as highly customized research and communication processes.

Surrogate Endpoints

The choice of an outcome is seemingly easy and apparent. For example, mortality is the dominant concern for many situations, and is seldom a difficult outcome to ascertain, unless there is a high loss to follow-up, which should not be a problem if the study is designed properly. However, if all cause death is the outcome this principal holds, if the determination is the specific reason for death, it becomes decidedly more difficult. This difficulty is because many deaths occur either outside the hospital where one has to rely on death certificates as the cause of death, or in hospital, where many patients have multi-organ disease, and trying to parse the specific cause is likely to be difficult. And yet, ascertaining the cause of death is essential for classifying disease-specific mortality in clinical research studies. As mentioned, death certificates often serve as the source of this information with the recognition that the cause of death on the death certificate is often fraught with misclassification (in fact in some states in the US the cause of death is not even entered). The potential for bias from this misclassification, and the fact that obtaining death certificates can often be time consuming and labor intensive is problematic. As a result, many studies also use a proxy–reported statement to determine the cause of death. Halanych et al. [46], assessed the validity of proxy-reported causes of death

Table 3.10 Approximate sample size given the treatment effect and control group "outcome"

	Treatment effect			
Rate in control group (%)	10 %	20 %	30 %	50 %
2	100,000	25,000	10,000	3,000
10	65,000	15,000	6,000	2,000
50	2,100	518	225	80

in 336 participants of the REGARDS Study. Trained experts used study data, medical records, death certificates, and proxy reports to adjudicate deaths. Adjudicated cause of death had a higher rate of agreement with proxy reports (73 %; Cohens kappa = .69) then with death certificates (63 % kappa = .54). Using the adjudicator cause of death as the "gold standard", the sensitivity for proxy reports was 50–89 % (depending on the cause) and specificity; 94–98 %, compared to death certificates, sensitivity 31–81 %. They concluded: "in many settings, proxy reports may represent a better strategy for determining the cause of death than reliance on death certificates".

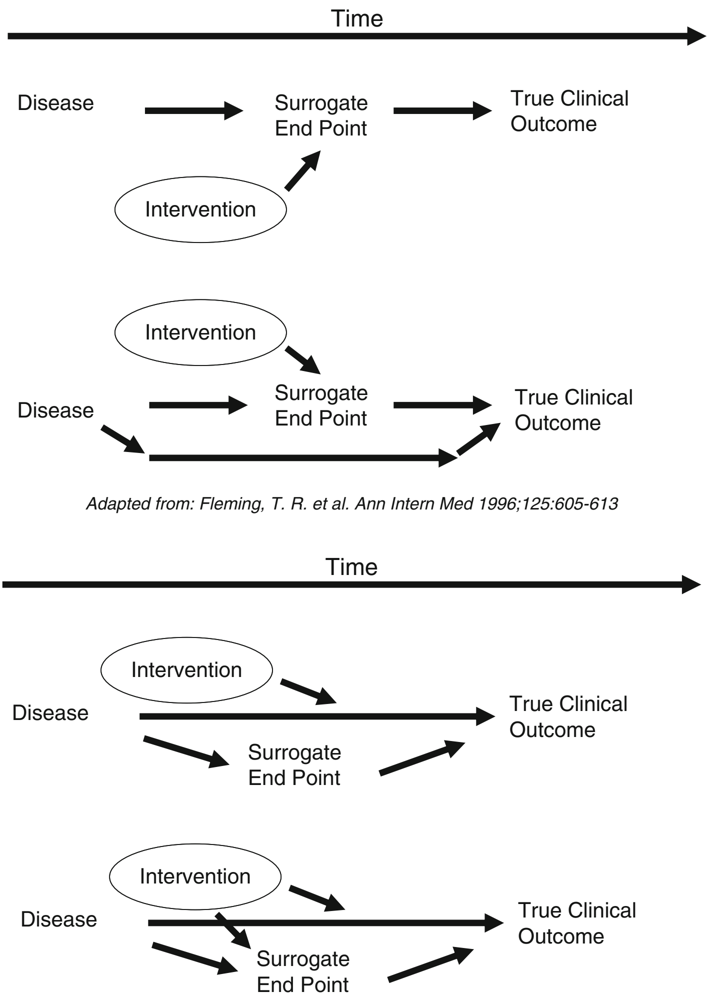
For many conditions mortality is not a frequent occurrence and only in the largest and longest trials would it be a practical choice. Thus, If the endpoints of interest are rare, RCTs have to be large (and expensive), so the question might arise as to how one can design a study to garner more endpoints? Several considerations for increasing endpoints include: extending the follow-up time, broaden the definition of an event, and, don't use the events of interest rather use surrogate endpoints. An example of this latter point might be a heart disease study in which coronary heart disease events or deaths (direct outcome of interest) and uses the surrogate of incident angina and/or revascularization procedures (this adds events) and even measures of atherosclerosis (moves to continuous measure). In a cancer study, one might be primarily interested in cancer recurrence and/or cancer death (direct), but one can move to the surrogate of tumor size that moves the outcome to a continuous measure.

In 1863, Farr said 'death is a fact, the rest is inference'. In choosing outcomes of interest, death or a disease event is usually the event of interest. However, as previously mentioned, it is frequently necessary to use a surrogate for the endpoint of interest, such as when the disease occurrence is rare and/or far in the future. The main variable that drives sample size and Study Power is the difference in the outcome between the intervention and the control group. Table 3.10 summarizes the sample size necessary based upon these aforementioned differences. One can see from Table 3.10 that most studies would have to be quite large unless the treatment difference is large, and for most outcomes these days, it is common to have treatment differences of no more than 20 %.

A surrogate endpoint is simply a laboratory value, sign, or symptom that is a substitute for the real outcome one is interested in [47]. The assumption is that changes induced in a surrogate endpoint accurately and nearly completely reflect changes in the clinically meaningful endpoint. To realize that assumption, an accurate well-documented model of the outcome of interest is a prerequisite, but it should be understood that the model is only that, and the model may be far from the truth. As is true of most definitions, there is debate about the best definition for a surrogate endpoint, and it is also important to distinguish surrogate endpoints from intermediate endpoints and statistical correlations. Speaking statistically, Prentice [48] has offered the following definition: '*a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint*.'

Examples of surrogate endpoints include blood pressure reduction in lieu of stroke (this has been termed a 'strong surrogate' by Anand et al.); [49] fasting blood sugar (or hemoglobin HbA1c) in lieu of diabetic complications; and bone mineral density in lieu of fractures. Surrogates are also commonly used early in drug development such as dose ranging or preliminary proof of efficacy ('developmental surrogates'). 'Supportive surrogates' are those outcomes that support and strengthen clinical trial data. The reasons for choosing a surrogate endpoint predominantly revolve around the fact that it might be easier to measure than the clinical endpoint of interest, or that it occurs early in the natural history of the disease of interest (and thus long-term trials are avoided). But as is true of almost any decision one makes in conducting a clinical trial, there are assumptions and compromises one has to make when choosing a surrogate endpoint. For example, many surrogates have been inadequately validated, and many if not most surrogates have several effect pathways (see Fig. 3.5). Other considerations for using a surrogate endpoint are that it should be easier to assess than the corresponding clinical endpoint, and in general, be more frequent; and, that an estimate of the expected clinical benefit should be derivable from the interventions effect upon the surrogate. An example of the controversy regarding surrogate endpoints is highlighted by the discussion of Kelsen [50] regarding the use of tumor regression as an adequate surrogate for new drugs to treat colorectal cancer. On the basis of a meta-analysis, Buyse et al. [51] proposed that surrogate endpoints of efficacy, without direct demonstration of an improvement in survival, could be used to identify effective new agents. The FDA, however, requires that there be a survival advantage before it approves such a drug. That is, a response rate higher than standard therapy (defined as tumor regression >50 %) is by itself an inadequate benefit for drug approval. As stated in the commentary by Kelsen '*the critical question in the debate over the adequacy of response rate as a surrogate endpoint for survival is whether an objective response to treatment is merely associated with a better survival, or whether the tumor regression itself lengthens survival*.'

There are differences in an intermediate endpoint, correlate, and a surrogate endpoint, although an intermediate endpoint may serve as a surrogate. Examples of intermediate endpoints include such things as angina pectoris, or hyperglycemic symptoms i.e. these are not the ultimate outcome of interest (MI, or death etc) but are of value to the patient should they be benefited by an intervention. Another example is from the earlier CHF literature where exercise-walking time was used as an intermediate endpoint as well as a surrogate marker for survival. A number of drugs improved exercise-walking time in the CHF patient; but long-term studies proved that the same agents that improved walking time actually resulted in earlier death. A hypothetical example of a surrogate 'misadventure' is exemplified by a



Adapted from: Fleming, T. R. et al. Ann Intern Med 1996;125:605-613

Fig. 3.5 Different surrogate paradigms

scenario where a new drug is used in pneumonia, and it is found to lower the patients white blood count (this used as a surrogate marker for improvement in the patients pneumonia). Subsequently, this hypothetical 'new drug' is found to be cytotoxic to white blood cells but obviously had little effect on the pneumonia. But, perhaps the most glaring example of a surrogate 'misadventure' is represented by a real trial – the Cardiac Arrhythmia Suppression Trial (CAST) [52]. At the time of CAST, premature ventricular contractions (PVC's) were thought to be a good surrogate for ventricular tachycardia or ventricular fibrillation, and thereby for sudden cardiac death (SCD). It was determined that many anti-arrhythmic agents available at the time or being developed reduced PVC's, and it was assumed would benefit the real outcome of interest, SCD. CAST was proposed to test the hypothesis that these

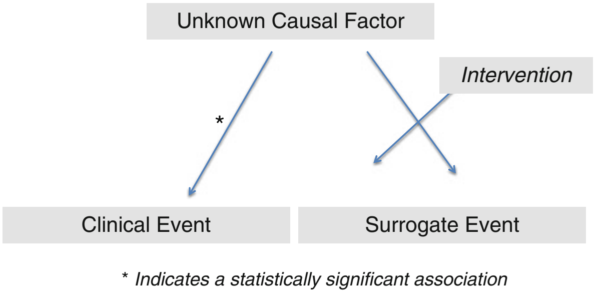


Fig. 3.6 Depicts a correlation (statistically significant) between a causal factor and a clinical event. While treatment impacted the surrogate event, it had no effect on the clinical event since it does not lie in the direct pathway

anti-arrhythmic agents did actually reduce SCD (in a post MI population) and this study was surrounded with some furor about the studies ethics, since a placebo control was part of the study design (it was felt strongly by many that the study was unethical since it was so likely that reduction in PVCs led to a reduction in SCD and how could one therefore justify a placebo arm). In fact, it turned out that the anti-arrhythmic therapy not only failed to reduce SCD, but in some cases it increased its frequency. A final example of surrogate misadventure occurred in 2007, when the Chairman of the FDA Advisory panel that reviewed the safety of rosiglitazone stated that the time has come to abandon surrogate endpoints for the approval of type 2 diabetes drugs. This resulted from the use of glycated hemoglobin as a surrogate for diabetes morbidity and mortality as exemplified in the ADOPT (A Diabetes Outcome Prevention Trial) study where patients taking rosiglitazone had a greater decrease in glycosylated hemoglobin than in patients taking comparator drugs, yet the risks of CHF and cardiovascular ischemia were higher with rosiglitazone [53].

Correlates may or may not be good surrogates. Recall, ‘that a surrogate endpoint requires that the effect of the intervention on the surrogate end-point predicts the effect on the clinical outcome-a much stronger condition than correlation.’ [47] Another major point of confusion is that between statistical correlation and proof of causality as demonstrated in Fig. 3.6 as discussed by Boissel et al. [54].

In summary, it should be understood that most (many) potential surrogates markers used in clinical research have been inadequately validated and that the surrogate marker must fully (or nearly so) capture the effect of the intervention on the clinical outcome of interest. However, many if not most treatments have several effect pathways and this may not be realized, particularly early in the research of a given intervention. Table 3.11 summarizes some of the issues that support using a surrogate. Surrogate endpoints are most useful in phase 1 and 2 trials where ‘proof of concept’ or dose-response is being evaluated. One very important additional down-side to the use of surrogate measures is a result of its effect on the safety evaluation of an intervention i.e. the ability to use smaller sample sizes and shorter trials imparted by the use of a surrogate endpoint, in order to gain insight into the benefit of an intervention results in the loss of important safety information.

Table 3.11 Support for and against the use of surrogate outcomes

Support for/against surrogates		
Factor	Favors surrogate	Does not favor surrogate
Biologic plausibility	Epidemiologic evidence extensive; excellent animal models pathogenesis and MOA understood; surrogate is late in causal pathway	Less extensive evidence; no animal model; MOA not understood, surrogate early in causal pathway
Success in clinical trials	Effect on surrogate has predicted outcome with other drugs in class and in disease	Inconsistent results across classes
Risk/benefit	Serious or life-threatening illness and no alternative treatment; large safety database; short term use; difficulty studying clinical endpoint	Less serious disease; little safety data; long term use; easy to study clinical endpoint

MOA mechanism of action

Selection of Endpoints

Table 3.10 makes the point that for most clinical trials, one of the key considerations is the difference in events between the investigational therapy and the control. It is this difference (along with the frequency of events) that drives the sample size and power of the study. From Table 3.10, one can compare the rate in the control group compared to the intervention effect. Thus, if the rate in the control group of the event of interest is high (say 20 %) and the treatment effect is 20 % (i.e. an expected 50 % reduction compared to control), a sample size of 266 patients would be necessary. Compare that to a control rate of 2 % and a treatment effect of 10 % (i.e. a reduction compared to control from 2 to 1.8 %), where a sample size of 97959 would be necessary. Often the question is asked; "What is a meaningful difference in endpoints?"

> A difference to be a difference must make a difference *(Gertrude Stein)*.

Primary and Secondary Endpoints

O'Neil [55] defines an endpoint as "*results, condition or events associated with individual study patients that are used to assess study treatments*". The characteristics of endpoint measures should include those that are easy to diagnose, easy to identify (i.e. no evaluator judgment needed), free of measurement error, reliable with repeated measures, have high internal validity and be directly linked to property of interest, and have good external validity.

Endpoints can be primary, secondary, tertiary, etc. A primary endpoint for a drug in development is a "clinical endpoint that provides evidence sufficient to fully categorize clinically the effect of a treatment that would support a regulatory claim for the treatment". A secondary endpoint is when there is "additional clinical characterization of a treatment but could not, by itself, be convincing of a clinically significant treatment effect". Tertiary and other endpoints are mostly exploratory. Some questions about secondary endpoints include:

- How does one interpret secondary endpoints when the primary endpoint for which the clinical trial was initially designed does not meet the proposed effect.
- Some argue for caution in making inferences from secondary endpoints, and certainly there are limitations and greater concerns for a secondary endpoint effect that is derived from only one study. The likelihood of replication of the finding in another study of identical size and design as a useful concept to guide this interpretation.
- O'Neill R. (1997) argues that "secondary endpoints ***cannot*** be validly analyzed if the primary endpoint does not demonstrate clear statistical significance" [55], while Davis, C.E. (1997) argues that "secondary endpoints ***can*** be validly analyzed, even if the primary endpoint does not provide clear statistical significance" [55].

In practice, it is rare that trials use a single endpoint, and endpoints frequently cover clinical events, symptoms, physiologic measures, quality of life etc. One example is taken from the "Multiple Sclerosis literature where the result of interest was neurological disability and endpoints included episodes" of focal neurological signs and symptoms, disability rating scales, MRI changes, and CSF changes.

Ultimately the choice of endpoints is a critical and challenging study design decision, based upon considerations such as the phase of development of the clinical question, the specific disease under study, the characteristics of the measure, and the questions the investigator wants answered by the trial. General guidelines in the choice of endpoints include the use of "hard endpoints" whenever possible ("hard" endpoints are clinical landmarks that are well-defined in the study protocol, are definitive with respect to disease process, and not subjective). It is true that some endpoints are useful and reliable even when they require some subjectivity, and the key issue is not the classification of an endpoint as "hard" or "soft", but how prone to measurement error the endpoint is.

Finally other arguments centered on study endpoints are that many advocate having a single primary endpoint, since this is what "drives" sample size calculations; and, multiple endpoints introduces the possibility of Type I error.

Composite Endpoints

It is generally realized that there is an increasing challenge to conduct adequately powered clinical trials. Most trials are designed to assess the time to some first event between two arms of a study. More and more frequently, different clinical events related to the target disease are combined to form a composite endpoint. Composite endpoints (rather than a single endpoint) are being increasingly used as effect sizes for most new interventions are becoming smaller. Effect sizes are becoming smaller because newer therapies need to be assessed when added to all clinically accepted therapies; and, thus the chance for an incremental change is reduced. For example, when the first therapies for heart failure were introduced, they were basically added to diuretics and digitalis. Now, a new therapy for heart failure would have to show benefit in patients already receiving more powerful diuretics, digitalis, angiotensin converting enzyme inhibitors and/or angiotensin receptor blockers, appropriately used beta adrenergic blocking agents, statins etc. To increase the ‘yield’ of events, composite endpoints are utilized (a group of individual endpoints that together form a ‘single’ endpoint for that trial). Thus, the rationale for composite endpoints comes from three basic considerations: statistical issues (sample size considerations due to the need for high event rates in the trial in order to keep the trial relatively small, of shorter duration and with less expense), the pathophysiology of the disease process being studied, and the increasing need to evaluate an overall clinical benefit. There are several downsides associated with the use of composite endpoints, one is that the benefits ascribed to an intervention are assumed to relate to all the components of the composite. Consider the example of a composite endpoint that includes death, MI, and urgent revascularization. In choosing the components of the composite, one should not be driven by the least important variable just because it happens to be the most frequent (e.g. death, MI, urgent revascularization, would be a problem if revascularization turned out to be the main positive finding). Another downside is that the first event within a composite endpoint may not reflect the most clinically important endpoint, and if the study is designed for time to first event, subsequent events within the composite will be missed. Thus incorporating subsequent events is seemingly rational [56]. Montori et al. provided guidelines for interpreting composite endpoints which included asking whether the individual components of composite endpoints were of similar importance, occurred with about the same frequency, had similar relative risk reductions, and had similar biologic mechanisms [57]. Armstrong and Westerhaut added to this by recommending that a strategy for future trials would be to include not just the initial event, but all events and report both per patient and overall rates; and, including a gradation of event severity (e.g. a large MI with heart failure has a very different meaning than a small periprocedural MI or a hemorrhagic stroke vs. a transient left arm weakness).

Freemantle et al. assessed the incidence and quality of reporting of composite endpoints in randomized trials and asked whether composite endpoints provide for greater precision but at the expense of greater uncertainty [58]. Their conclusion was that the reporting of composite outcomes is generally inadequate and as a result, they provided several recommendations regarding the use of composite endpoints such as following the CONSORT guidelines, interpreting the composite endpoint rather than parsing the individual endpoints, and defining the individual components of the composite as secondary outcomes. The reasons for their recommendations stemmed from their observations that in many reports they felt that there was inappropriate attribution of the treatment effects on specific endpoints when only composite endpoints yielded significant results, the effect of dilution when individual endpoints might not all react in the same direction, and the effect of excessively influential endpoints that are not associated with irreversible harm.

Table 3.12 An example of using MACE as a composite endpoint

Acute vs. non acute MI	MACE definition
1.7 (1.2–2.4)	Death; MI; stent thrombosis
1.15 (0.98–1.6)	Death; MI; stent thrombosis; target vessel revascularization
1.13 (0.95–1.4)	Death; MI; stent thrombosis; repeat revascularization
Multi-lesion vs. one lesion attempt	
1.1 (0.75–1.5)	Death; MI; stent thrombosis
1.35 (1.2–1.75)	Death; MI; stent thrombosis; target vessel revascularization
1.25 (0.01–1.52)	Death; MI; stent thrombosis; repeat revascularization

Adapted from: Kip et al. [60]

In an accompanying editorial by Lauer and Topel they list a number of key questions that should be considered when composite endpoints are reported or when an investigator is contemplating their use [59]. First, is whether the end points themselves are of clinical interest to patients and physicians, or are they surrogates; second, how nonfatal endpoints are measured (e.g. is judgment involved in the end point ascertainment, or is it a hard end point); third, how many individual endpoints make up the composite and how are they reported (ideally each component of the composite should be of equal clinical importance – in fact, this is rarely the case); and finally, how are non fatal events analyzed – that is are they subject to competing risks. As they point out, patients who die cannot later experience a non fatal event so a treatment that increases the risk of death may appear to reduce the risk of nonfatal events, and vice versa [59].

Kip et al. [60] reviewed the problems with the use of composite endpoints in cardiovascular studies. The term "major adverse cardiac events:" or MACE is used frequently in cardiovascular studies, a term that was born with the percutaneous coronary intervention studies in the 1990s. Kip et al. noted that MACE encompassed a variety of composite endpoints, the varying definitions of which could lead to different results and conclusions, leading them to the recommendation that MACE as a composite endpoint should be avoided. Table 3.12 from their article demonstrates this latter point rather well.

As mentioned above, composite endpoints are commonly used to increase event rates in an effort to increase statistical power. However, attention towards whether the individual components of the composite are likely to be differentially affected by the intervention is important. Bethel et al. performed a meta-analysis to determine the effect of angiotensin-converting enzyme inhibitors or angiotensin receptor blockers on individual cardiovascular outcomes; and then applied these treatment effects to two different composite cardiovascular endpoints. They found that although composite endpoints did augment event rates, they did not necessarily increase statistical power, and in fact, in some cases reduced it [61]. As they noted, *"occurrence of the composite endpoint must be in keeping with the duration and intensity of follow-up within a clinical trial and should reflect prior knowledge of* *the magnitude of expected treatment benefits. If insufficient data exist to estimate the treatment effect, pooled data based on plausibly similar mechanisms of action may be used instead."*

Central to the selection of endpoints is how the endpoints are adjudicated, and for most large clinical trials this is generally accomplished with a centralized system. This is most important when the primary endpoint is a nonfatal event since the definition may be somewhat subjective. The main concern relative to adjudication is to avoid differential misclassification-that is to adjudicate events that are biased by applying the outcome definition variably or by knowing to which treatment assignment the patient was in (as might occur in an open-label study). The idea is that with a central adjudication system in which the adjudicators are blinded as to the treatment assignment and apply the same definitions uniformly, will yield the least biased assessment. However, this aforementioned concept has not been adequately investigated. Granger et al. reviewed the literature concerning the rationale and justification for central adjudication, and came to the conclusion that it has not been shown to improve the ability to determine treatment effects, and may be overly complex and overused. And yet, the FDA and the scientific community derive confidence in the validity of results when central adjudication is performed [62].

Trial Duration

A critical decision in performing or reading about a RCT (or any study for that matter) is the specified duration of follow-up, and how that might influence a meaningful outcome. Many examples and potential problems exist in the literature, but basically in interpreting the results of any study (positive or negative) the question should be asked 'what would have happened had a longer follow-up period been chosen?' An example is the Canadian Implantable Defibrillator Study (CIDS) [63]. CIDS was a RCT comparing the effects of defibrillator implantation to amiodarone in preventing recurrent sudden cardiac death in 659 patients. At the end of study (a mean of 5 months) a 20 % relative risk reduction occurred in all-cause mortality, and a 33 % reduction occurred in arrhythmic mortality, when ICD therapy was compared with amiodarone (this latter reduction did not reach statistical significance). At one center, it was decided to continue the follow-up for an additional mean of 5.6 years in 120 patients who remained on their originally assigned intervention [64]. All-cause mortality was increased in the amiodarone group. The Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) trial is an example of a potential problem in which study duration could have been problematic (but probably wasn't) [65]. The central hypothesis of MIRACL was that early rapid and profound cholesterol lowering therapy with atorvastatin could reduce early recurrent ischemic events in patients with unstable angina or acute non-Q wave infarction. Often with acute intervention studies, the primary outcome is assessed at 30 days after the sentinel event. From Fig. 3.7 one can see that there was no difference in the primary outcome at 30 days. Fortunately the study specified a 16-week follow-up, and a significant difference was seen at that time point. Had the

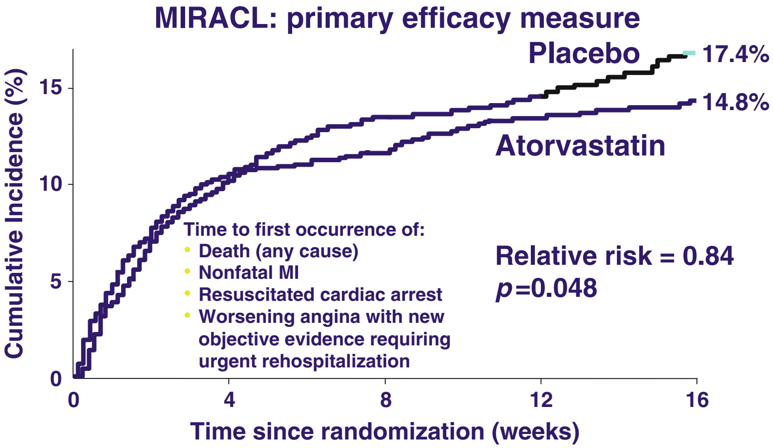


Fig. 3.7 The results of MIRACL for the primary outcome. What would have been the conclusion for the intervention if the pre-specified study endpoint was 1 month? (Adapted from Schwartz et al. [65])

study been stopped at 30 days the ultimate benefit would not have been realized. Finally, an example from the often cited controversial ALLHAT study which demonstrated a greater incidence in new diabetes in the diuretic arm as assessed at the study end of 5 years [66]. The investigators pointed out that this increase in diabetes did not result in a statistically significant difference in adverse outcomes when the diuretic arm was compared to the other treatment arms. Many experts have subsequently opined that the trial duration was too short to assess adverse outcomes from diabetes, and had the study gone on longer that it is likely that a significant difference in adverse complications from diabetes would have occurred.

The Devil Lies in the Interpretation

It is interesting to consider and important to reemphasize, that intelligent people can look at the same data and render differing interpretations. MRFIT is exemplary of this principal, in that it demonstrates how mis-interpretation can have far-reaching effects. One of the conclusions from MRFIT was that reduction in cigarette smoking and cholesterol was effective, but '*possibly an unfavorable response to antihypertensive drug therapy in certain but not all hypertensive subjects*' led to mixed benefits [31]. This 'possibly unfavorable response' (thought to be due to diuretic based hypokalemia) has since been at least questioned if not proven to be false.

Differences in interpretation was also seen in the alpha-tocopherol, beta carotene cancer study [33]. To explain the lack of benefit and potential worsening of cancer risk in the treated patients, the authors opined that perhaps the wrong dose was used, or that the intervention period was to short, since '*no known or described mechanisms and no evidence of serious toxic effects of this substance* (beta carotene) *in humans*' had been observed. This points out how ones personal bias can influence ones 'shaping' of the interpretation of a trials results. Finally, there are many examples of trials where an interpretation of the results is initially presented only to find that after publication differing interpretations are rendered. Just consider the recent controversy over the interpretation of the ALLHAT results [66].

Causal Inference, and the role of **the Media** in reporting clinical research will be discussed in chapters 16 and 20.

Conclusions

While randomized clinical trials are the 'gold standard' clinical research design, there remains many aspects of trial design that must be considered before accepting the studies results, even when the study design is a RCT. Starzi et al. in their article entitled 'Randomized Trialomania? The Multicentre Liver Transplant Trials of Tacrolimus' outline many of the roadblocks and pitfalls that can befall even the most conscientious clinical investigator [67]. Ioannidis presents an even more somber view of clinical trials, and has stated 'there is increasing concern that in modern research', false findings may be the majority or even the vast majority of published research claims. He points out that this should not be surprising since it can be proven that most (one can argue many if not most) claimed research findings are false [68]. Also, many feel that misleading interpretations result from an over-reliance on statistical testing, that is, that the strength of evidence is often judged by conventional tests that rely heavily on statistical significance, with less attention paid to the clinical significance or practical importance of treatment effects [69]. Kaul and Diamond cite three particular technical limitations to the interpretation of the results from a clinical trial: the emphasis of statistical significance over clinical importance, the use of composite endpoints, and the use of subgroup analyses (refer to sections on composite endpoints and subgroup analysis above). Relative to the over-reliance on statistical testing is the controversy that surrounds relying on the p value, and as a wit opined *"a p value is no substitute for a brain"* (anonymous source cited in Kaul and Diamond). The significance level that is used most commonly is the P value ≤ 0.05 that represents the maximum probability that is tolerated for rejecting a hypothesis that is in fact true. But in contrast to the p ≤ 0.05 standard for statistical significance is that there are no guidelines for what difference is clinically significant and some then equate the two. Kaul and Diamond conclude that "while statistical significance tells us whether a difference is likely to be real, it does not place that reality into meaningful clinical context by telling us the difference is small, large, trivial, or important. A formal evaluation of clinical importance (using frequentist confidence intervals, the number needed to treat and the number needed to harm, or Bayesian probabilities), given the overall risk-benefit-cost profile of each therapeutic intervention, should be included in the analysis, interpretation, and presentation of the results of clinical trials." Table 3.13 provides a list of at least 12 misconceptions about P values [70].

Table 3.13 Twelve P-value misconceptions

1. If $P = .05$, the null hypothesis has only a 5 % chance of being true
2. A nonsignificant difference (e.g., $P > .05$) means there is no difference between groups
3. A statistically significant finding is clinically important
4. Studies with P values on opposite sides of .05 are conflicting
5. Studies with the same P value provide the same evidence against the null hypothesis
6. $P \leq .05$ means that we have observed data that would occur only 5 % of the time under the null hypothesis
7. $P \leq .05$ and $P < .05$ mean the same thing
8. P values are properly written as inequalities (e.g., "$P < .02$" when $P = .015$)
9. $P \leq .05$ means that if you reject the null hypothesis, the probability of a type I error is only 5 %
10. With a $P \leq .05$ threshold for significance, the chance of a type I error will be 5 %
11. You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible
12. A scientific conclusion or treatment policy should be based on whether or not the P value is significant

Adapted from Goodman [70]

One final note of caution revolves around the use of reading or reporting only abstracts in decision-making. As Toma et al. noted, 'not all research presented at scientific meetings is subsequently published, and even when it is, there may be inconsistencies between these results and what is ultimately printed' [71]. They compared RCT abstracts presented at the American College of Cardiology sessions between 1999 and 2002, and subsequent full-length publications. Depending upon the type of presentation (e.g. late breaking trials vs. other trials) 69–79 % were ultimately published; and, discrepancies between meeting abstracts and publication results were common even for the late breaking trials (see Chap. 19 for further discussion of abstracts) [71].

References

1. Glasser SP, Howard G. Clinical trial design issues: at least 10 things you should look for in clinical trials. J Clin Pharmacol. 2006;46:1106–15.
2. Grady D, Herrington D, Bittner V, Blumenthal R, Davidson M, Hlatky M, et al. Cardiovascular disease outcomes during 6.8 years of hormone therapy: Heart and Estrogen/progestin Replacement Study follow-up (HERS II). JAMA. 2002;288:49–57.
3. Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. JAMA. 1998;280:605–13.
4. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. JAMA. 2002;288:321–33.
5. Grady D, Rubin SM, Petitti DB, Fox CS, Black D, Ettinger B, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med. 1992;117:1016–37.

6. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. Prev Med. 1991;20:47–63.
7. Sullivan JM, Vander Zwaag R, Hughes JP, Maddock V, Kroetz FW, Ramanathan KB, et al. Estrogen replacement and coronary artery disease. Effect on survival in postmenopausal women. Arch Intern Med. 1990;150:2557–62.
8. Bhatt DL, Cavender MA. Are all clinical trial sites created equal? J Am Coll Cardiol. 2013;61:580–1. doi:10.1016/j.jacc.2012.10.024.
9. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367:1355–60. PMC3771340.
10. Butler J, Subacius H, Vaduganathan M, Fonarow GC, Ambrosy AP, Konstam MA, et al. Relationship between clinical trial site enrollment with participant characteristics, protocol completion, and outcomes: insights from the EVEREST (Efficacy of Vasopressin Antagonism in Heart Failure: Outcome Study with Tolvaptan) trial. J Am Coll Cardiol. 2013;61:571–9. doi:10.1016/j.jacc.2012.10.025.
11. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002;359:57–61.
12. Loscalzo J. Clinical trials in cardiovascular medicine in an era of marginal benefit, bias, and hyperbole. Circulation. 2005;112:3026–9.
13. Bienenfeld L, Frishman W, Glasser SP. The placebo effect in cardiovascular disease. Am Heart J. 1996;132:1207–21.
14. Clark PI, Leaverton PE. Scientific and ethical issues in the use of placebo controls in clinical trials. Annu Rev Public Health. 1994;15:19–38.
15. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. N Engl J Med. 1994;331:394–8.
16. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA. 2005;294:2203–9.
17. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ. 1948;ii:769–82.
18. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
19. Gum PA, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. JAMA. 2001;286:1187–94.
20. Reviews of statistical and economic books, Student's Collected Papers. J R Stat Soc. 1943;106:278–9.
21. Fleming TR. Addressing missing data in clinical trials. Ann Intern Med. 2010;154:113–7. PMC3319761.
22. A Village of 100 In. 2nd ATS Media ed: A Step Ahead.
23. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Engl J Med. 1991;325:445–53.
24. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Endarterectomy for asymptomatic carotid artery stenosis. JAMA. 1995;273:1421–8.
25. Lang JM. The use of a run-in to enhance compliance. Stat Med. 1990;9:87–93; discussion –5.
26. Franciosa JA. Commentary on the use of run-in periods in clinical trials. Am J Cardiol. 1999;83:942–4. A9.
27. Pablos-Mendez A, Barr RG, Shea S. Run-in periods in randomized trials: implications for the application of results in clinical practice. JAMA. 1998;279:222–5.
28. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet. 2002;359:696–700.
29. Shem S. The house of god. In: Palgrave Macmillan; 1978:280.
30. Smith DH, Neutel JM, Lacourciere Y, Kempthorne-Rawson J. Prospective, randomized, open-label, blinded-endpoint (PROBE) designed trials yield the same results as double-blind, placebo-controlled trials with respect to ABPM measurements. J Hypertens. 2003;21:1291–8.

31. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial. Risk factor changes and mortality results. JAMA. 1982;248:1465–77.
32. Mayo E. The human problems of an industrial civilization. New York: Macmillan; 1993.
33. Alpha-Tocopherol T, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N Engl J Med. 1994;330:1029–35.
34. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319:670–4.
35. The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. N Engl J Med. 1980;303:1038–41.
36. The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. N Engl J Med. 1980;302:250–6.
37. Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. N Engl J Med. 1979;301:1410–2.
38. Howard G, Chambless LE, Kronmal RA. Assessing differences in clinical trials comparing surgical vs nonsurgical therapy: using common (statistical) sense. JAMA. 1997;278:1432–6.
39. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000;355:1064–9.
40. Sleight P. Debate: subgroup analyses in clinical trials: fun to look at – but don't believe them! Curr Control Trials Cardiovasc Med. 2000;1:25–7.
41. Amarenco P, Goldstein LB, Szarek M, Sillesen H, Rudolph AE, Callahan 3rd A, et al. Effects of intense low-density lipoprotein cholesterol reduction in patients with stroke or transient ischemic attack: the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) trial. Stroke. 2007;38:3198–204.
42. Black HR, Elliott WJ, Grandits G, Grambsch P, Lucente T, White WB, et al. Principal results of the Controlled Onset Verapamil Investigation of Cardiovascular End Points (CONVINCE) trial. JAMA. 2003;289:2073–82.
43. Weir MR, Ferdinand KC, Flack JM, Jamerson KA, Daley W, Zelenkofske S. A noninferiority comparison of valsartan/hydrochlorothiazide combination versus amlodipine in black hypertensives. Hypertension. 2005;46:508–13.
44. Kaul S, Diamond GA, Weintraub WS. Trials and tribulations of non-inferiority: the ximelagatran experience. J Am Coll Cardiol. 2005;46:1986–95.
45. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. JAMA. 2006;295:1147–51.
46. Halanych JH, Shuaib F, Parmar G, Tanikella R, Howard VJ, Roth DL, et al. Agreement on cause of death between proxies, death certificates, and clinician adjudicators in the Reasons for Geographic and Racial Differences in Stroke (REGARDS) study. Am J Epidemiol. 2011;173:1319–26. PMC3101067.
47. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125:605–13.
48. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. 1989;8:431–40.
49. Anand IS, Florea VG, Fisher L. Surrogate end points in heart failure. J Am Coll Cardiol. 2002;39:1414–21.
50. Kelsen DP. Surrogate endpoints in assessment of new drugs in colorectal cancer. Lancet. 2000;356:353–4.
51. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Meta-Analysis Group in Cancer. Lancet. 2000;356:373–8.
52. Greene HL, Roden DM, Katz RJ, Woosley RL, Salerno DM, Henthorn RW. The Cardiac Arrhythmia Suppression Trial: first CAST… then CAST-II. J Am Coll Cardiol. 1992;19:894–8.
53. FDA Adviser Questions Surrogate Endpoints for Diabetes Drug Approvals. In: Medpage Today; 2007.
54. Boissel JP, Collet JP, Moleur P, Haugh M. Surrogate endpoints: a basis for a rational approach. Eur J Clin Pharmacol. 1992;43:235–44.

55. O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. Control Clin Trials. 1997;18:550–6. discussion 61–7.
56. Armstrong PW, Westerhout CM. The power of more than one. Circulation. 2013;127:665–7. doi:10.1161/CIRCULATIONAHA.112.000627.
57. Montori VM, Busse JW, Permanyer-Miralda G, Ferreira I, Guyatt GH. How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: should I dump this lump? ACP J Club. 2005;143:A8.
58. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003;289:2554–9.
59. Lauer MS, Topol EJ. Clinical trials – multiple treatments, multiple end points, and multiple lessons. JAMA. 2003;289:2575–7.
60. Kip KE, Hollabaugh K, Marroquin OC, Williams DO. The problem with composite end points in cardiovascular studies. J Am Coll Cardiol. 2008;51:701–7. doi:10.1016/j.jacc.2007.10.034.
61. Bethel MA, Holman R, Haffner SM, Califf RM, Huntsman-Labed A, Hua TA, et al. Determining the most appropriate components for a composite clinical trial outcome. Am Heart J. 2008;156:633–40. doi:10.1016/j.ahj.2008.05.018.
62. Granger CB, Vogel V, Cummings SR, Held P, Fiedorek F, Lawrence M, et al. Do we need to adjudicate major clinical events? Clin Trials. 2008;5:56–60. doi:10.1177/1740774507087972.
63. Connolly SJ, Gent M, Roberts RS, Dorian P, Roy D, Sheldon RS, et al. Canadian implantable defibrillator study (CIDS): a randomized trial of the implantable cardioverter defibrillator against amiodarone. Circulation. 2000;101:1297–302.
64. Bokhari F, Newman D, Greene M, Korley V, Mangat I, Dorian P. Long-term comparison of the implantable cardioverter defibrillator versus amiodarone: eleven-year follow-up of a subset of patients in the Canadian Implantable Defibrillator Study (CIDS). Circulation. 2004;110:112–6.
65. Schwartz GG, Olsson AG, Ezekowitz MD, Ganz P, Oliver MF, Waters D, et al. Effects of atorvastatin on early recurrent ischemic events in acute coronary syndromes: the MIRACL study: a randomized controlled trial. JAMA. 2001;285:1711–8.
66. The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the antihypertensive and lipid lowering treatment to prevent heart attack trial (ALLHAT). JAMA 2002;288:2981–97.
67. Starzl TE, Donner A, Eliasziw M, Stitt L, Meier P, Fung JJ, et al. Randomised trialomania? The multicentre liver transplant trials of tacrolimus. Lancet. 1995;346:1346–50.
68. Ioannidis JPA. Why most published research findings are false. PLoS. 2005;2:696–701.
69. Kaul S, Diamond GA. Trial and error. How to avoid commonly encountered limitations of published clinical trials. J Am Coll Cardiol. 2010;55:415–27. doi:10.1016/j.jacc.2009.06.065.
70. Goodman SA. A dirty dozen: Twelve P-value misconceptions. Semin Hematol. 2008;45:135–40. doi:10.1053/j.seminhematol.2008.04.003.
71. Toma M, McAlister FA, Bialy L, Adams D, Vandermeer B, Armstrong PW. Transition from meeting abstract to full-length journal article for randomized controlled trials. JAMA. 2006;295:1281–7.

Chapter 4
Alternative Interventional Study Designs

Stephen P. Glasser

A man who does not habitually wonder is but a pair of spectacles behind which there is no eye

(Thomas Carlyle) [1]

Abstract There are many variations to the classical randomized controlled trial. These variations are utilized when, for a variety of reasons, the classical randomized controlled trial would be impossible, inappropriate, or impractical. Some of the variations are described in this chapter and include: equivalence and non-inferiority trials; crossover trials; N of 1 trials, case-crossover trials, and externally controlled trials. Large simple trials, and prospective randomized, open-label, blinded endpoint trials are discussed in another chapter.

Keywords Equivalence/noninferiority testing • Superiority testing • PROBE design • Factorial design • Assay sensitivity • Consistency assumption • N of 1 trial • Crossover design • Case-crossover design • Adaptive design • Registry randomized control trial • Null hypothesis

There are a number of variations of the ‘classical’ RCT design. For instance, many view the classical RCT as having an exposure group compared to a placebo control group, using a parallel design, and a 1:1 randomization scheme. However, in a given RCT, there may be several exposure groups (e.g. utilizing different doses of the drug under study), and the comparator group may be an active control rather than a

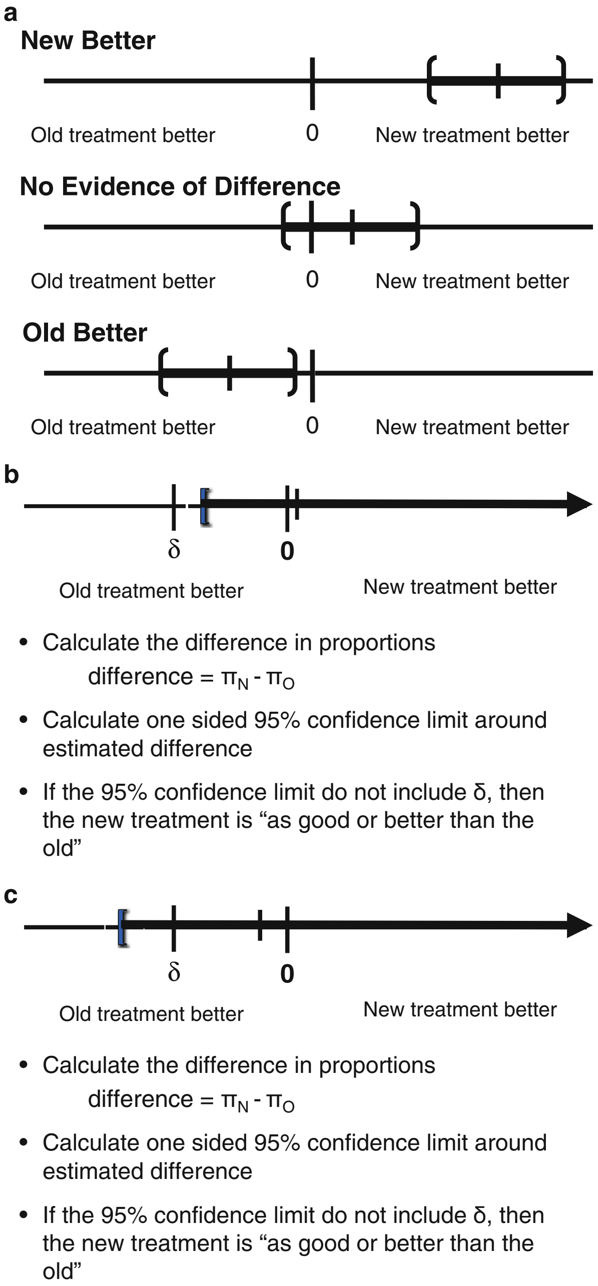
S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_4,

placebo control; and, some studies may have both. By an active control, it is meant that the control group receives an already approved intervention. For example, a new anti-hypertensive drug could be compared to placebo or could be compared to a drug already approved by the FDA and used in the community (frequently, in this case, the manufacturer of the investigational drug will compare their drug to the currently most frequently prescribed drug for the indication of interest). The decisions regarding the use of a comparator are based upon a number of considerations and discussed more fully under the topic entitled equivalence testing. Also, the randomization sequence may not be 1:1, particularly if (for several reasons, ethical issues may be one example) one wanted to reduce the number of subjects exposed to placebo. Also, rather than parallel groups, there may be a titration schema built into the design. On occasion, the study design could incorporate a placebo withdrawal period in which at the end of the double blind comparison, the intervention group is subsequently placed on placebo (this can be done single-blind or double-blind). In this latter case, retesting 1 or 2 weeks later occurs with comparison to the original placebo group. Other common variants to the classical RCT are discussed in more detail below.

Traditional Versus Equivalence/Non-inferiority Testing (See Tables 3.6 in Chap. 3 and 4.1 in This Chapter)

As discussed in Chap. 3, most clinical trials have been designed to assess if there is a difference in the efficacy to two (or more) alternative treatment approaches (with placebo ideally being the comparator treatment). Consider the fact that for evidence of efficacy there are two distinct approaches: to demonstrate a difference-showing superiority of the investigational drug to control (placebo, active, lower dose) which then demonstrates the drug effect; or, to show equivalence or non-inferiority to an active control (i.e. the investigational drug is of equal efficacy or not worse than an active control). That is, one can attempt to demonstrate that there is similarity to a known effective therapy (active control) and attributing the efficacy of the active control drug to the investigational drug, thereby demonstrating a drug effect (i.e. equivalence). Since nothing is perfectly equivalent, equivalence means within a margin predetermined by the investigator (termed the equivalence margin). Non-inferiority trials on the other hand aim to demonstrate that the investigational drug is not worse than the control, but once again by a defined amount (i.e. not worse by a given amount – the non-inferiority margin), the margin (M or δ) being that amount no larger than the effect the active control would be expected to have in the study. As will be discussed later, this margin is not easy to determine and requires clinical judgment; and, this represents one of the limitations of these kinds of trials [2]. These aforementioned approaches are presented diagrammatically in Fig. 4.1a–c.



- Calculate the difference in proportions
  difference = $\pi_N - \pi_O$
- Calculate one sided 95% confidence limit around estimated difference
- If the 95% confidence limit do not include δ, then the new treatment is "as good or better than the old"

- Calculate the difference in proportions
  difference = $\pi_N - \pi_O$
- Calculate one sided 95% confidence limit around estimated difference
- If the 95% confidence limit do not include δ, then the new treatment is "as good or better than the old"

Fig. 4.1 (**a**) Outcomes for traditional (superiority) testing. (**b**) Outcomes for equivalence testing. Since the lower confidence bound is not beyond theta, the null has not been rejected. (**c**) Outcome in equivalence testing. Since the lower confidence bound is beyond theta, the null is rejected and therefore the two treatments are "equivalent"

As also discussed in Chap. 3, there are a number of reasons for the increased interest in equivalence and non-inferiority trials including the ethical issues associated with placebo controls. In general, for studies of efficacy, placebo-controls are preferable to active controls, due to the placebo's ability to distinguish an effective treatment from a less effective treatment. The ethical issues surrounding the use of a placebo-control aside, there are other issues that have led to the increasing interest and use of equivalence and non-inferiority studies. For example, clinical trials are increasingly being required to show benefits on clinical endpoints rather than on surrogate endpoints at the same time that the incremental benefit of new treatments is getting smaller. This has led to the need for larger, longer, and more costly trials; and, this has resulted in the need to design trials that are less expensive. Additional issues are raised by the use of equivalence/non-inferiority trials, such as assay sensitivity, the aforementioned limitations of defining the margins, and the constancy assumption.

Assay Sensitivity

Assay sensitivity is a property of a clinical trial defined as the ability of the trial to distinguish effective from ineffective treatments [3]. That is, assay sensitivity is the ability of a specific clinical trial to demonstrate a treatment difference if such a difference truly exists [3]. Assay sensitivity depends on the effect size one needs to detect. One, therefore, needs to know the effect of the control drug in order to determine the trials assay sensitivity. There is then an inherent, usually unstated, assumption in an equivalence/non-inferiority trial, namely that the active control is similarly effective in the particular study one is performing (i.e., that one's trial has assay sensitivity), compared to a prior study that utilized a placebo comparator. However, this aforementioned assumption is not necessarily true for all effective drugs, is not directly testable in the data collected (if there is no placebo group to serve as an internal standard); and this, in essence, causes an active control equivalence study to have elements of a historically controlled study [4].

A trial that demonstrates superiority has inherently demonstrated assay sensitivity; but, a trial that finds the treatments to be similar, cannot distinguish (based upon the data alone) between a true finding, and a poorly executed trial that just failed to show a difference. Thus, an equivalence/non-inferiority trial must rely on the assumption of assay sensitivity, based upon quality control procedures and the reputation of the investigator. The International Conference on Harmonization (ICH) guidelines (see Chap. 6) list a number of factors that can reduce assay sensitivity, and includes: poor compliance, poor diagnostic criteria, excessive measurement variability, and biased endpoint assessment [5]. Thus, assay sensitivity can be more directly ascertained in an active control trial only if there is an 'internal standard,' a control vs. placebo comparison as well as the control vs. test drug comparison (e.g. a three-arm study).

Advantages of the Equivalence/Non-inferiority Approach

As discussed above, the application of equivalence testing permits a definitive statement that the new treatment is 'as good' (if the null hypothesis is rejected), and depending upon the circumstances, this statement may meet the needs of the manufacturer, who may only want to make the statement that the new treatment is as good as the established treatment, with the implication that the new treatment is preferred because it may require less frequent dosing, or be associated with fewer side effects, less invasiveness etc. On the other hand, the advantage of superiority testing is that one can definitively state if one treatment is better (or worse) than the other, with the downside that if there is not evidence of a difference, you cannot state that the treatments are the same (recall, that the null hypothesis is never 'accepted' – it is simply a case where it cannot be rejected, i.e. 'there is not sufficient evidence in these data to establish if a difference exists').

Disadvantages or Limitations of Equivalence/Non-inferiority Studies

The disadvantages of equivalence/non-inferiority testing include: (1) that the choice of the margin chosen to define whether two treatments are equivalent is difficult; (2) that it requires clinical judgment and should have clinical relevance (variables that are difficult to measure); (3) the assumption that the control would have been superior to placebo (assumed assay sensitivity) had a placebo had been employed (constancy assumption- that is, one expects the same benefit in the equivalence/non-inferiority trial as occurred in a prior placebo controlled trial); and, (4) having to determine the margin such that it is not greater than the smallest effect size (that of the active drug vs. placebo) in prior placebo controlled trials [6]. In addition, there is some argument as to whether the analytic approach in equivalence/non-inferiority trials should be ITT or Per Protocol (Compliers Only) [7]. While ITT is recognized as valid for superiority trials, the inclusion of data from patients not completing the study in equivalence/non-inferiority trials, could bias the results towards the treatments being the same, which could then result in an inferior treatment appearing to be non-inferior or equivalent. On the other hand, using the compliers only (per protocol) analysis may bias the results in either direction. Most experts in the field argue that the Per Protocol (some like to say non ITT analysis implying that it is as close to ITT analysis as possible) analysis is preferred for equivalence/non-inferiority trials but some still argue for the ITT approach [7]. Also, blinding does not protect against bias as much in equivalence/non-inferiority trials as it does with superiority trials-since the investigator, knowing that the trial is assessing equality may subconsciously assign similar ratings to the treatment responses of all patients.

Table 4.1 Approaches to hypothesis testing in clinical trials

RCT Hypothesis testing		
Hypothesis	Superiority	Equivalence/noninferiority
Null	New = Old	New < Old ± margin
Alternative	New > Old	New = Old
Null rejected	New is different than Old	New is at least as effective as Old
Failure to reject the null	Did not show that New is different that Old	Did not show that New is as effective as Old

The Null Hypothesis in Equivalence/Non-inferiority Trials (Table 4.1)

> It is a beautiful thing, the destruction of words…Take 'good' for instance, if you have a word like 'good' than is there need for the word "bad"? 'Ungood' will do just as well [8]

Recall that with traditional (superiority) hypothesis testing, the null hypothesis states that 'there is no difference between treatment groups' (i.e. New = Established, or placebo). Rejecting the null, then allows one to definitively state if one treatment is better than another (i.e. New > or < Established). The disadvantage is if at the conclusion of an RCT there is not evidence of a difference, one cannot state that the treatments are the same, or as good as one to the other.

Equivalence/non-inferiority testing in essence 'flips' the traditional null and alternative hypotheses. Using this approach, the null hypothesis is that the new treatment is worse than the established treatment (i.e. New < Old); that is, rather than assuming that there is no difference, the null hypothesis in equivalence/non-inferiority trials is that a difference exists and the new treatment is inferior. Some distinguish between equivalence and noninferiority, since strictly speaking equivalence means that the treatment effect is between the + and – margins and is therefore 2-sided, while noninferiority implies that the new treatment is "no worse than the old treatment and therefore is 1-sided. However, many in the field and an extension of the CONSORT Statement [9] suggest that two-sided confidence intervals are appropriate for most noninferiority trials, so the need for separating the two approaches is questionable.

Just as in traditional testing, the two actions available resulting from statistical testing is: (1) reject the null hypothesis, or (2) failure to reject the null hypothesis. However, with noninferiority/equivalence testing, rejecting the null hypothesis is making the statement that the new treatment is not worse than established treatment, implying the alternative, that is, that the new treatment is as good as (i.e. New ≧ Established). Hence, this approach allows a definitive conclusion that the new treatment is at least as good, or is not inferior to the established.

As mentioned before, a caveat is the definition of 'as good as,' which is defined as being in the 'neighborhood' or having a difference that is so small as to be considered clinically unimportant (generally, event rates within ±2 % – this is known as the equivalence or non-inferiority margin usually indicted by the symbol δ). The need for this 'neighborhood' that is considered 'as good as' exposes the first shortcoming of equivalence/non-inferiority testing – having to make a statement that "I reject the null hypothesis that the new treatment is worse than the established, and accept the alternative hypothesis that it is as good *and by that I mean that it is within at least X % of the established*" (the wording in italics are rarely included in the conclusions of a manuscript). A second caveat of equivalence/non-inferiority testing is that no definitive statement can be made that there is evidence that the new treatment is better or worse. Just as in traditional testing, one never accepts the null hypothesis – one only fails to reject it. Hence if the null is not rejected, all one can really say is that there is no evidence in these data that the new treatment is as good as or better than the old treatment. In equivalence trials, the conventional significance testing has little relevance, since failure to detect a difference does not imply equivalence. Rather, results should be reported with point estimates and confidence limits with the equivalence margin kept in mind.

In summary**,** the design of equivalence trials should mirror that of earlier successful trials of the active comparator as closely as possible [10] and, analysis strategies should not center on intention-to-treat (since ITT tends to reduce the difference between the intervention and control, it biases towards equivalence). Jones et al. also discuss why equivalence trials generally need to be larger than their placebo controlled counterparts, and why the standard of conduct needs to be especially high in terms of withdrawals, losses, and protocol deviations.

A potential concern has been raised over the rapid growth of noninferiority trials. For example, If novel therapy "A" is non-inferior to existing therapy "B" which itself was brought to market based upon non-inferiority data compared to therapy "C", the non-inferiority margin becomes more difficult to ascertain. Some potential ways one can overcome this is by comparing A to C directly but this may not be feasible if B has supplanted C in clinical practice. Alternatively, the margin for comparing A to B can be set to narrow limits, but this will increase the sample size.

One might ask; which is the 'correct' approach, superiority or equivalence testing? There is simply no general answer to this question; rather, the answer depends on the major goal of the study. But, once an approach is taken, the decision cannot be changed in post-hoc analysis. That is, the format of the hypotheses has to be tailored to the major aims of the study and must then be followed. An example of one innovative study in which the design combined a non-inferiority and superiority analysis is the Rivaroxaban versus Warfarin in Nonvalvular Atrial Fibrillation (ROCKET AF Study) which was a double-blind phase 3 study in more than 14,000 patients with atrial fibrillation. Patients were randomized to 20-mg rivaroxaban once daily (or 15 mg in patients with moderate renal impairment at screening) or to dose-adjusted warfarin (titrated to an international normalized ratio [INR] of 2.5). In the ROCKET-AF trial, patients were randomly assigned to receive either rivaroxaban or warfarin. In a per protocol, as-treated analysis, rivaroxaban was found to be noninferior to warfarin with respect to the primary end point of stroke or systemic embolism. As a pivotal trial for the new oral factor Xa inhibitor, rivaroxaban met its primary end point showing the drug was noninferior to warfarin. Disappointingly, however, in the same study the intention-to-treat superiority analysis failed to show the drug had an advantage, statistically, over warfarin. In an on-treatment analysis addressing the superiority question, however, rivaroxaban fared better, the rates of the composite major and non-major clinically relevant bleeding were comparable in the rivaroxaban- and warfarin-treatment arms [11].

Crossover Design

In crossover designs, both treatments (investigational and control) are administered sequentially to all subjects, and randomization occurs in terms of which treatment each patient receives first. In this manner each patient serves as his/her own control. The two treatments can be an experimental drug vs. placebo or an experimental drug compared to an active control. The value of this approach beyond being able to use each subject as their own control, centers on the ability (in general) to use smaller sample sizes. For example, a study that might require 100 patients in a parallel group design might require fewer patients in a crossover design. But like any decision made in clinical research there is always a 'price to pay.' For example, the washout time between the two treatments is arbitrary, and one has to assume that they have eliminated the likelihood of carryover effects from the first treatment period (plasma levels of the drug in question are usually used to determine the duration of the crossover period, but in some cases the tissue level of the drug is more important). Additionally, there is some disagreement as to which baseline period measurement, (the first baseline period or the second baseline period-they are almost always not the same) should be used to compare the second period effects.

N of 1 Trials

During a clinical encounter, the benefits and harms of a particular treatment are paramount; and, it is important to determine if a specific treatment is benefiting the patient or if a side effect is the result of that treatment. This is particularly a problem if adequate trials have not been performed regarding that treatment. Inherent to any study is the consideration of why a patient might improve as a result of an intervention. Of course, what is generally hoped for is that the improvement is the result of the intervention. However, improvement can also be a result of the disease's natural history, placebo effect, or regression to the mean (see Chap. 7). Clinically (in a practice setting), a response to a specific treatment is assessed by a trial of therapy, but this is usually performed without rigorous methodological standards so the results may be in question; and, this has led to the n of 1 trial (sometimes referred to as an RCT crossover study in single patients). In its usual form, n of 1 trials are randomized, double-blind, multiple crossover comparisons of an active drug against placebo in individual patients, and may be useful for determining individual treatment effects and as a tool to estimate heterogeneity of treatment effects in a population. An example of heterogeneity of treatment effects is the study by Pedro-Botet et al. [12]. Whereas the mean percent LDL-C response following 12 months of atorvastatin therapy (10 mg qd) was in the order of 35 %, the heterogeneity of effect is nicely portrayed in Fig. 4.2.

The requirements of the n of 1 design are: the patient receives active, investigational therapy during one period, and alternative therapy (e.g. placebo) during another

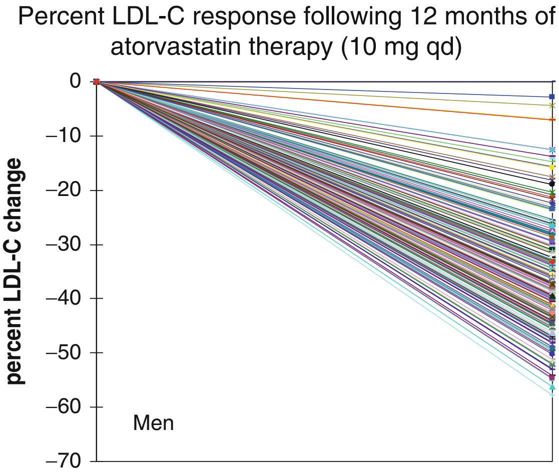


Fig. 4.2 The percent in LDL-C lowering in response to 12 months of Atorvastatin Therapy (10 mg/QD) (Pedro-Botet et al. [12])

period as would occur with typical crossover designs. As is also true of crossover designs, the order of treatment from one patient to another is randomly varied, and other attributes-blinding/masking, ethical issues, etc.- are adhered to just as they are in the classical RCT. In contrast to the typical crossover design however, at a pre-specified point (perhaps a given number of crossovers, or degree of improvement or deterioration) the patient's involvement in the study is stopped and their response held until all patients complete the trial.

There are at least three obvious sources of variability in clinical trials. Firstly, pure differences occur between patients: e.g. some are more seriously ill than others. Secondly, there is variability within patients: even given the same treatment they, or their measurements, may vary from time to time. Thirdly, some patients may react more favorably to a given treatment than other patients. The parallel group trial does not and cannot distinguish between types of variability; and, while the standard crossover trial will distinguish between the first type of variability and the other two it does not distinguish easily between the second and third. The n of 1 trial does address some of these issues in variability.

The n-of-1 trial does have some characteristics of the "playing the winner, dropping the loser" adaptive design (see below), but unlike this latter design, the patient in the n-of-1 trial may end the study (for that patient) when a pre-specified endpoint is reached. Some caveats to consider before designing an n-of-1 trial is that these trials are oriented towards symptomatic treatments that have rapid improvement upon treatment initiation, and rapid loss of efficacy upon therapy discontinuation. The use of this trial design is thus problematic when dealing with chronic disease therapies in which the acute response does not predict long term outcome, when the anticipated treatment effect is difficult to differentiate from random fluctuations of disease, and when treatment effects are small (i.e. hard to detect in an individual patient).

Table 4.2 Possible outcomes and stopping rules in N of 1 trials

Result	Continue	Stop
Benefit likely, harm unlikely	×	
Benefit possible, harm unlikely	×	
Benefit possible, harm possible		×
Benefit unlikely, harm unlikely		×
Benefit possible, harm possible		×
Inconclusive result		×

Adapted from: Mahon et al. [13]

An example of the n of 1 trial was reported by Mahon et al. [13] regarding the evaluation of the efficacy of theophylline for irreversible chronic airflow limitation. As these authors state; "*though the efficacy of theophylline for irreversible chronic airflow limitation has been established in conventional randomized controlled trials, its efficacy in individual patients is often in doubt.*" Patients fulfilling the entry criteria for this trial (n = 31), were randomized by coin toss to either an n of 1 trial or standard treatment by a person unaware of their baseline characteristics. Some patients entered an open trial of theophylline that was given for 2 weeks at their previously used dose, and all patients were uncertain that theophylline was helpful while taking it openly. This was established by the patient not affirmatively answering the question, "Are you certain that theophylline is helping you?" Each patient was then randomized to a double-blind, multiple crossover comparison of theophylline vs. placebo and their results were compared to use of theophylline as standard therapy (administered according to published guidelines). For the n of 1 trial participant's, the order of theophylline and placebo were randomly determined and the physician monitoring the response was blinded as to treatment assignment. If deterioration occurred the patient was immediately switched to the other treatment, while if on the other hand the deterioration occurred during the second treatment period they were switched back to the first period treatment. In this way this study design of early switching or stopping treatment is designed to limit the ethical problem of a patient remaining symptomatic during alternate (particularly placebo) treatment.

Potentially, a number of different scenarios could occur as outlined in Table 4.2. The difference in theophylline use at 6 months between the n of 1 trial and standard practice groups – without significant changes in exercise capacity and quality of life – suggests that the suspected bias of standard practice towards unnecessary treatment is real, by virtue of the much greater use of theophylline among standard practice patients (difference 47 %).

In 2011, Gabler et al. [14] reviewed 108 n of 1 trials done between 1985 and 2010 on 2,154 participants, and concluded that n of 1 trails are a useful tool for enhancing therapeutic precision in a wide range of conditions, and should be conducted more often.

Factorial Designs

Many times it is possible to evaluate 2 or even 3 treatment regimens in one study. In the Physicians Health Study, for example, the effect of aspirin and beta carotene was assessed [15]. Aspirin was being evaluated for its ameliorating effect on myocardial

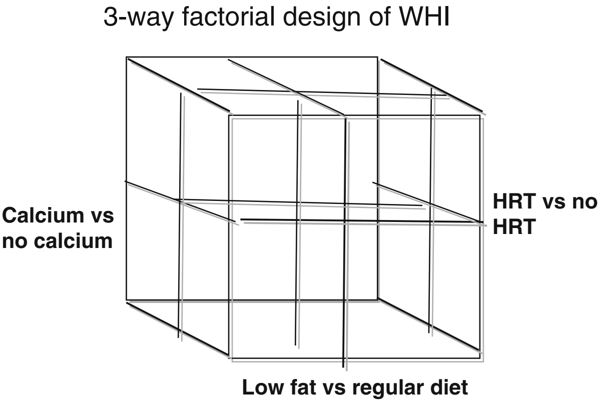


Fig. 4.3 Three-way Latin square design

infarction, and beta carotene on cancer. Subjects were randomized to 1 of 4 groups; placebo and placebo, aspirin and placebo, beta carotene and placebo, and aspirin plus beta carotene. In this manner, each drug could be compared to placebo, and any interaction of the two drugs in combination could also be evaluated. This type of design certainly can add to the efficiency of a trial, but this is counterbalanced by increased complexity in performing and interpreting the trial results. In addition, the overall trial sample size is increased (4 randomized groups instead of the usual 2), but the overall sample size is likely to be less than the total of two separate studies, one addressing the effect of aspirin and the other of beta carotene. In addition two separate studies would lose the ability to evaluate treatment interactions, if that is a concern. Irrespective, costs (if it is necessary to answer both questions) should be less with a factorial design compared to two separate studies, since recruitment, overhead etc. should be less. The Woman's Health Initiative is an example of a three-way factorial design [16]. In this study, hormone replacement therapy, calcium/vitamin D supplementation, and low fat diets were evaluated (see Fig. 4.3). Overall, factorial designs can be seductive but can be problematic, and it is best used for unrelated research questions, both as it applies to the intervention as well as the outcomes.

Case-Crossover Design

Case-crossover designs are a variant, having components of a crossover, and a case–control design. The case cross over design was first introduced by Maclure in 1991 [17]. It is usually applied to study transient effects of brief exposures on the occurrence of a 'rare' acute onset disease. The presumption is that if there are precipitating events preceding the outcome of interest, these events should be more frequent during the period immediately preceding the outcome, than at a similar period that is more distant from the outcome. For example, if physical and/or mental stress triggers sudden cardiac death (SCD), one should find that SCD occurred more frequently during or shortly after these stressors. In a sense, it is a way of assessing whether the patient was doing anything unusual just before the outcome of interest. As mentioned above, case-crossover studies are related to a prospective crossover design in that each subject passes through both the exposure (in the case-crossover design this is called the hazard period) and ‘placebo’ (the control period). The case-cross over design is also related to a case–control study in that it identifies cases and then looks back for the exposure (but in contrast to typical case–control studies, in the case-crossover design the patient serves as their own control). Of course, one needs to take into account the times when the exposure occurs but is not followed by an event (this is called the exposure-effect period). The hazard period is defined empirically (one of this designs limitations, since this length of time may be critical yet somewhat arbitrary) as the time period before the event (say an hour or 30 min) and is the same time given to the exposure-effect period. A classic example of the case-crossover design was reported by Hallqvist et al., where the triggering of an MI by physical activity was assessed [18]. To study possible triggering of first events of acute myocardial infarction by heavy physical exertion, Halqvist et al. conducted a case-crossover analysis. Interviews were carried out in 699 myocardial infarction patients. The relative risk from vigorous exertion was 6.1 (95 % confidence interval: 4.2, 9.0), while the rate difference was 1.5 per million person-hours [18].

In review, the strengths of the case-crossover study design include using subjects as their own control (self matching decreases between-person confounding, although if certain characteristics change over time there can be individual confounding), and improved efficiency (since one is analyzing relatively rare events). In the example of the Halqvist study, although MI is common, MI just after physical exertion is not [18]. Weaknesses of the study design, besides the empirically determined time for the hazard period, include: recall bias, and that the design can only be applied when the time lag between exposure and outcome is brief and the exposure is not associated with a significant carryover effect.

Externally Controlled Trials (Before-After Trials)

Using historical control’s as a comparator to the intervention is problematic, since the natural history of the disease may have changed over time, and certainly sample populations may have changed (e.g. greater incidence of obesity, more health awareness, new therapies, etc. now vs. the past). However, when an RCT with a concomitant control cannot be used (this can occur for a variety of reasons-see example below) there is a way to use a historical control that is not quite as problematic. Olson and Fontanarosa cite a study by Cobb et al. to address survival during out of hospital ventricular fibrillation [19]. The study design included a pre-intervention period (the historical control) during which emergency medical technicians (EMT) administered defibrillation as soon as possible after arriving on scene of a patient in cardiac arrest. This was followed by an intervention period where the EMT performed CPR for 90 s before defibrillation. In this way many of the problems of typical historical controls can be overcome in that in the externally controlled design, one can use the same sites and populations in the ‘control’ and intervention groups as would be true of a typical RCT, it is just that the control is not concomitant. Another example is that of Sipilä et al. who assessed the impact of a guideline implementation intervention on antihypertensive drug prescribing; specifically, to assess the effects of a multifaceted (education, audit, and feedback, local care pathway) quality program. The proportions of patients receiving specific antihypertensive drugs and multiple antihypertensive drugs were measured before and after the intervention for three subgroups of hypertension patients: hypertension only, with coronary heart disease, and with diabetes.

Nonconventional Clinical Trial Designs

As the field of clinical trial methodology evolves, the need for alternative designs increases. This is reviewed by Howard [20] as it related to studies of stroke, but clearly it is not limited to that area. Howard outlined four such nonconventional approaches: dose selection trials; adaptive clinical trials; shift analysis; and Bayesian analysis.

Briefly, dose selection trials allow for dose adjustment as the trial proceeds primarily based upon the occurrence and frequency of any adverse effects at the dose being studied (unless the event rate is so low that it is not likely to be seen in a limited number of patients). The intent is to find the “optimal” dose (i.e. the highest potential dose that is associated with a low occurrence of adverse drug events). Adaptive clinical trials refers to a study design that is adjusted based upon data collected initially (sometimes confused with group- sequential studies) [21]. As Howard noted, “*Shih eloquently relates that group sequential methodology has the goal of saving lives or resources, whereas the adaptive clinical trial approach has the goal of saving the study*” [21].

“Shift analysis” allows for a reduction in sample size or gain in power, but further discussion is beyond the scope of this book. Bayesian analysis (Also see Chap. 14) is a potentially rapidly rising approach in clinical trials. Simplifying, the characteristic that defines any statistical approach is how it deals with uncertainty (see Chap. 18). The traditional approach to dealing with uncertainty is the frequentist approach, which deals with fixed sample sizes based upon prior data; but otherwise the information present from prior studies is not incorporated into the study being now implemented. That is, with the frequentist approach “the difference between treatment groups is assumed to be an unknown and fixed parameter”. A Bayesian approach uses previous data to develop a prior distribution of potential differences between treatment groups and updates this data with that collected during the trial being performed to develop a posterior distribution (this is akin to the discussion in Chap. 14 that addresses pre and post test probability).

There are strong advocates of the frequentist and the Bayesian approach, which should indicate that neither is perfect and that one or the other may be preferable in certain situations. The argument then devolves to not which is better, but in which circumstance might one be preferable. Further discussion is also beyond this books scope, but should be of interest to the more advanced student.

Adaptive Designs

It is recognized that increased spending on biomedical research has not increased success rates of drug development-due to: diminished margin for improvement, chronic diseases are harder to study, rapidly escalating costs, and pharmaceutical company mergers that have decreased new-drug candidates. This has led to more innovative designs for evaluating drug efficacy. Adaptive designs give flexibility for identifying the optimal clinical benefit of a test treatment without ***"significantly"*** undermining the validity and integrity of the intended study. Some examples are the use of adaptive randomization; group sequential analysis (discussed in Chap. 9), and sample size re-estimation. Adaptive designs can be prospective (e.g. adaptive randomization, stopping a trail early due to safety, futility, or efficacy, dropping the loser (playing the winner); concurrent (e.g. modifying inclusion/exclusion criteria, modifying a dose/regimen and treatment duration); or, retrospective (e.g. changes in the statistical plan prior to database lock or unblinding of treatment codes). Whereas some adaptive changes require no or little statistical adjustment (e.g. dropping a treatment arm, modifying dosing paradigms, modifying randomization ratios; modifying subject selection, modifying visit schedules, or modifying study eligibility criteria), some do (e.g. requiring statistical adjustments, resizing a study, and allowing for the inclusion of subjects who participated in earlier drug development studies in a later development study – although this not generally recommended). What generally cannot be recommended in adaptive designs are: changes in the primary endpoint, and more than 1 adjustment to sample size.

One example of an adaptive design and to be contrasted to n of 1 trials (see above) is "playing the winner – dropping the loser", (this is an example of adaptive randomization). This design allows for dropping inferior treatment responses and adding additional arms, so it is useful, for example, in early drug development studies when there are uncertainties regarding dose levels. An example of how this is operationalized is starting out with a probability of 50 % randomization to both groups (allocation ratio of 1:1), and you randomize a patient to one of the treatments. If they do well, you increase the likelihood that the next randomization will be to the same group, the basic idea is to keep adjusting the likelihood randomization to a specific treatment group in order to increase the chances of the beneficial treatment going to the winner. For example, choose a staring base, say 20 subjects, and a 1:1 randomization scheme (10 A/20 A + B). One then randomizes a patient, and one assumes that they went to "A" and did well. Then the likelihood of the next patient being randomized to A would change from 50:50 to 52:48. Let's assume that the next patient despite increased odds of going to A in fact gets randomized to B and does poorly. We now further increase the likelihood going to "A" (say to 55 %). If a patient is randomized to B and does well, one adjusts the chances that the next patient will be randomized to B, and so on. Over time, if one group is doing better the likelihood of a patient being randomized to that group increases.

Registry-Based Randomized Clinical Trials (RRCT)

The Thrombus Aspiration in ST Elevation Myocardial Infarction in Scandinavia Trial (TASTE) was reported in 2013 and the study design caused a lot of excitement [22]. The TASTE trial enrolled ST elevation MIs as they entered a long-standing Swedish Web System registry for another goal. Based upon that registry (which provided comprehensive data collection and follow-up, TASTE built a web-based randomization that allocated 7,200 patients to either treatment by thrombus aspiration followed by PTCA or PCTA only. The enthusiasm about the design was that it allowed for completeness of follow-up at a lower cost, and no commercial involvement. As Laure and D'Agostino point out, "*with this clever design, which leveraged clinical information that was already being gathered for the registry and for other preexisting databases, the investigators were able to quickly identify potential participants, to enroll thousands of patients in little time, to avoid filling out long case report forms, to obtain accurate follow-up with minimal effort, and to report their findings, all for less than a typical RO1 grant*" [23]. They do go on to point out a number of potential problems with the RRCT, however, including the quality of the data, missing data, privacy, blinding etc. But, the RRCT potentially presents an alternative to the standard RCT in countries with large observational registry programs.

References

1. Breslin JE. Quote Me. Ontario: Hounslow Press; 1990.
2. Siegel JP. Equivalence and noninferiority trials. Am Heart J. 2000;139:S166–70.
3. Assay Sensitivity. In: Wikipedia. Available from: http://en.wikipedia.org/wiki/Assay_sensitivity
4. Snapinn SM. Noninferiority trials. Curr Control Trials Cardiovasc Med. 2000;1:19–21.
5. Walter SD. Choice of effect measure for epidemiological data. J Clin Epidemiol. 2000;53:931–9.
6. D'Agostino Sr RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. Stat Med. 2003;22:169–86.
7. Wiens BL, Zhao W. The role of intention to treat in analysis of noninferiority studies. Clin Trials. 2007;4:286–91.
8. Diamond GA, Kaul S. An Orwellian discourse on the meaning and measurement of noninferiority. Am J Cardiol. 2007;99:284–7.
9. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA. 2006;295:1152–60.
10. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ. 1996;313:36–9.
11. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med. 2011;365:883–91. doi:10.1056/NEJMoa1009638.
12. Pedro-Botet J, Schaefer EJ, Bakker-Arkema RG, Black DM, Stein EM, Corella D, et al. Apolipoprotein E genotype affects plasma lipid response to atorvastatin in a gender specific manner. Atherosclerosis. 2001;158:183–93.
13. Mahon J, Laupacis A, Donner A, Wood T. Randomised study of n of 1 trials versus standard practice. BMJ. 1996;312:1069–74.

14. Gabler NB, Duan N, Vohra S, Kravitz RL. N-of-1 trials in the medical literature: a systematic review. Med Care. 2011;49:761–8. doi:10.1097/MLR.0b013e318215d90d.
15. Hennekens CH, Eberlein K. A randomized trial of aspirin and beta-carotene among U.S. physicians. Prev Med. 1985;14:165–8.
16. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. JAMA. 2002;288:321–33.
17. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol. 1991;133:144–53.
18. Hallqvist J, Moller J, Ahlbom A, Diderichsen F, Reuterwall C, de Faire U. Does heavy physical exertion trigger myocardial infarction? A case-crossover analysis nested in a population-based case-referent study. Am J Epidemiol. 2000;151:459–67.
19. Olson CM, Fontanarosa PB. Advancing cardiac resuscitation: lessons from externally controlled trials. JAMA. 1999;281:1220–2.
20. Howard G. Nonconventional clinical trial designs: approaches to provide more precise estimates of treatment effects with a smaller sample size, but at a cost. Stroke. 2007;38:804–8.
21. Shih WJ. Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: a comparison. Stat Med. 2006;25:933–41.
22. Frobert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. N Engl J Med. 2013. [Epub ahead of print]. doi:10.1056/NEJMoa1308789.
23. Lauer MS, D'Agostino RB, Sr. The randomized registry trial – the next disruptive technology in clinical research? N Engl J Med. 2013. [Epub ahead of print]. doi:10.1056/NEJMp1310102.

Chapter 5
Phase 4 (Postmarketing) Research

Stephen P. Glasser, Elizabeth Delzell, and Maribel Salas

Research is what I'm doing when I don't know what I'm doing [2].

Wernher von Braun

Abstract Postmarketing research is a generic term used to describe all activities after the drug approval by the regulatory agency, such as the Food and Drug Administration (MedWatch: voluntary reporting by health professionals. http://www.fda.gov/medwatch/report/hcp.htm. Accessed 12 Oct 2006), European Medicines Agency (EMA), or other regulatory agencies. Postmarketing studies concentrate much more (but not exclusively) on safety and effectiveness, and they can contribute to the drugs implementation through labeling changes, length of the administrative process, pricing negotiations, and marketing. However, the fact that not all FDA mandated (classical phase IV trials) research consists of randomized controlled trials (RCTs), and not all postmarketing activities are limited to safety issues (pharmacovigilance), these terms require clarification. This chapter attempts

Less than 50 % of this chapter is taken from "Importance and challenges of studying marketed drugs: what is a phase IV study? Common clinical research designs, registries, and self-reporting systems" [1]. With permission of the publisher.

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

E. Delzell, Ph.D.
Department of Epidemiology, School of Public Health, University of Alabama
at Birmingham, Birmingham, AL, USA

M. Salas, M.D., D.Sc., M.Sc.
Medical Director, Worldwide Safety Strategy, Pfizer Inc, Collegeville, PA, USA

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_5,

to clarify the confusing terminology; and, to discuss many of the postmarketing research designs-both their place in clinical research as well as their limitations.

Keywords Postmarketing research • Surveillance study • Pharmacovigilance study • Phase 4 trial • Large simple trial • PROBE design • Effectiveness (pragmatic) trial • Drug utilization/pharmacoepidemiologic study • Comparative effectiveness research

In the past, postmarketing research, postmarketing surveillance and pharmacovigilance were synonymous with phase IV studies because the main activities of the regulatory agency (e.g. FDA) were focused on the monitoring of adverse drug events and inspections of drug manufacturing facilities and products [3]. However, the fact that not all FDA mandated (classical phase IV trials) research consists of randomized controlled trials (RCTs), and not all postmarketing activities are limited to safety issues (pharmacovigilance), these terms require clarification. Information from a variety of sources is used to establish the efficacy and short-term safety (<3 years) of medications used to treat a wide range of conditions. Premarketing studies (Table 5.1) consist of phase I-III trials, and are represented by pharmacokinetic and pharmacodynamic studies, dose ranging studies, and for phase III trials the gold standard randomized, placebo-controlled (or active controlled), double blind, trial (RCT). Approximately only 20 % of the drugs that enter phase I are approved for marketing [3]. RCTs remain the 'gold standard' for assessing the efficacy and to a lesser extent, the safety of new therapies; however, they do have significant limitations that promote caution in generalizing their results to routine clinical practice [2, 3]. For example, because of the strict inclusion and exclusion criteria mandated in most controlled studies, a limited number of patients who are relatively homogeneous are enrolled. Elderly patients, women, pregnant women, children, and those deemed not competent to provide informed consent are often excluded from such trials [4–7]. RCTs may also suffer from selection or volunteer bias. For example, clinical studies that include extended stays in a clinic may attract unemployed patients, and studies that involve a free physical examination may attract those concerned that they are ill. Studies that offer new treatments for a given disease may inadvertently select patients who are dissatisfied with their current therapy [7].

RCTs have other limitations as well. For example, the stringent restrictions regarding concomitant medications and fixed treatment strategies bear only modest resemblance to the ways in which patients are treated in actual practice [8, 9]. The aforementioned difference creates a situation dissimilar from routine clinical practice in which many or even most patients are taking multiple prescription and over-the-counter medications or supplements to manage both acute and chronic conditions [10, 11]. RCTs also generally include intensive medical follow-up in terms of number of medical visits, number and/or type of tests and monitoring events, that is usually not possible in routine clinical care [12]. Also, unintended adverse events (UAEs) are unlikely to be revealed during phase III trials since the usual sample sizes of such studies and even the entire New Drug Application (NDA) may range from hundreds to only a few thousand patients. For example, discovering an UAE

Table 5.1 Premarketing study designs for FDA approval

I. Phase I-III studies
a. Pharmacokinetic and pharmacodynamic studies
b. Dose-ranging studies
c. RCTs (Efficacy studies)
1. With or without crossover designs
2. Drug withdrawal designs
3. Placebo or active controls

Table 5.2 Estimated number or subjects (Study size) to find an adverse event of stated frequency

Frequency of ADE	Estimated number of subjects	Trial type
1 % (1/100)	1,000	Clinical trial
0.1 % (1/1,000)	10,000	Large clinical trial
0.01 % (1/10,000)	100,000	Post market surveillance
0.001 % (1/100,000)	1,000,000	Long term surveillance

with a frequency of 0.1 % would require a sample size of more than 10,000 participants (Table 5.2). Castle [13] further elaborated on this issue by asking the question 'how large a population of treated patients should be followed to have a good chance of picking up one, two, or three cases of an adverse reaction?' He notes that if one defines 'good chance' as a 95 % probability, one has to still factor in the expected incidence of a naturally occurring event that simulates the adverse event of interest. If one assumes no background incidence of such events, and the expected incidence of the adverse event of interest is 1 in 10,000, then by his assumptions, it would require 65,000 patients to pick up an excess of three adverse events.

Phase III trials also are not useful for detecting UAEs that occur only after exposure to long-term therapy, because of the insufficient length of follow-up time of the majority of phase III trials, nor do they provide information on long-term effectiveness and safety. All of the restrictions characteristic of controlled clinical studies may result in overestimation of the efficacy and underestimation of the potential for UAEs of the medication being evaluated [8, 12, 14]. As a result of these limitations, additional complementary approaches to evaluation of medication efficacy, effectiveness and safety are taking on increasing importance.

Postmarketing research (Table 5.3) is a generic term used to describe all activities after drug approval by the regulatory agency, such as the Food and Drug Administration [15]. Other regulatory agencies, such as the EMA, use the term post-authorization studies instead of post-marketing studies [16]. Postmarketing studies concentrate much more (but not exclusively) on safety and effectiveness, and they can contribute to the drugs implementation through labeling changes, length of the administrative process, pricing negotiations, and marketing. If post-authorization studies are focused on safety they are then known as post-authorization safety studies (PASS), and these can be interventional or non-interventional. According to the EMA, a PASS refers to "*any study relating to any authorized medicinal product conducted with the aim of identifying or quantifying a safety hazard, confirming the safety profile of the medicinal product, or measuring the effectiveness of risk management*" [17].

Table 5.3 Postmarketing study designs

I. FDA 'mandated or negotiated' studies (phase IV)
- a. Any study design may be requested including studies of
  - i. Drug-drug interactions
  - ii. Formulation advancement
  - iii. Special safety
  - iv. Special populations (e.g. post-authorization safety studies-PASS)
- b. 'phase V' trials

II. Non FDA 'mandated or negotiated' studies
- a. RCTs
  - i. Superiority vs. equivalence testing
  - ii. Large simple trials
  - iii. PROBE designs
  - iv. 'phase V' trials
- b. Surveillance studies
  - i. Pharmacovigilance studies
  - ii. Effectiveness studies
  - iii. Drug utilization studies
  - iv. Observational epidemiology studies

III. Health Services Research (HSR)

IV. Health Outcomes Research (HOR)

V. Implementation research

Note: We have not included a discussion of HSR or HOR in this review. Implementation Research will be discussed in Chap. 13

The most commonly used approaches for monitoring drug safety in humans are based on spontaneous reporting systems, automated linkage data, patient registries, case series and case reports, case–control studies, and data obtained directly from an interventional study. Since there are major limitations from relying on case reports or on voluntary reporting, postmarketing research has become an integral part of the drug evaluation process for assessing adverse events [18–23]. However, falling under the rubric of postmarking research is a wide variety of study designs and approaches, each with its own strengths and limitations. Postmarketing studies (Fig. 5.1 and Table 5.3) are not only represented by a much broader array of study designs, they have clearly differentiated goals compared to premarketing studies. Examples of study designs that might fall under the rubric of postmarketing research are phase IV clinical trials, practice-based clinical experience studies (mostly extinct now), large simple trials (LSTs), equivalence trials, post-marketing surveillance studies such as effectiveness studies, pharmacovigilance studies, PASS, and pharmacoeconomic studies.

There are several initiating mechanisms for postmarketing studies: (1) those required by a regulatory agency as a condition of the drug's approval (these are referred to as postmarketing commitments or PMC's); (2) those that are initiated by the pharmaceutical company to support various aspects of the development of that drug; (3) investigator initiated trials that may be as scientifically rigorous as

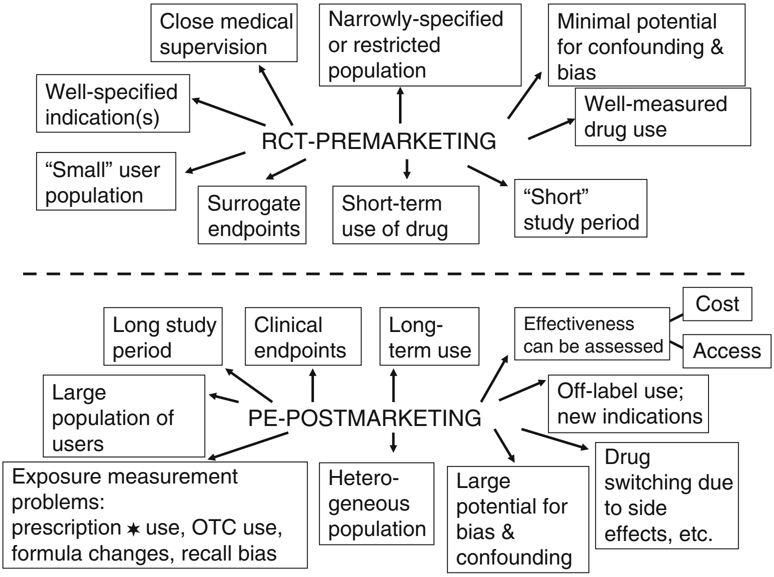

Close medical supervision
Narrowly-specified or restricted population
Minimal potential for confounding & bias
Well-specified indication(s)
Well-measured drug use
RCT-PREMARKETING
"Small" user population
Surrogate endpoints
Short-term use of drug
"Short" study period
Long study period
Clinical endpoints
Long-term use
Effectiveness can be assessed
Cost
Access
Large population of users
PE-POSTMARKETING
Off-label use; new indications
Exposure measurement problems: prescription ✶ use, OTC use, formula changes, recall bias
Hetero-geneous population
Large potential for bias & confounding
Drug switching due to side effects, etc.

Fig. 5.1 Contrasts between pre- and post-marketing studies

phase III RCTs, but occur after drug approval (a recent example is some of the Vioxx studies that ultimately questioned the drugs safety); and, (4) investigator initiated observational studies. The more scientifically rigorous postmarketing studies (particularly if they are RCTs) are sometime referred to as 'phase V' trials. This review will discuss each of the common types of postmarketing research studies and examples will be provided in order to highlight some of the strengths and limitations of each.

Sweeping regulatory changes are pushing Pharma to do more than has been expected of them in the past. Additional questions are being asked such as: How do drugs compare with other drugs marketed for the same indication (i.e. comparative effectiveness trials); and, how do drugs differ among subpopulations (e.g. elderly, pediatric patients, pregnant women, patients with renal or hepatic impairment etc.), and how does the new drug perform under real life conditions (i.e. effectiveness trials).

FDA 'Mandated or Negotiated' Studies (Phase IV Studies)

Phase IV studies are most often concerned with safety issues and usually have prospectively defined end-points aimed at answering safety questions. Any type of study (these include standard RCTs, observational studies, drug-drug interaction studies, special population studies, etc.-See Table 5.3) may be requested by the FDA upon NDA approval; and, these are frequently called Phase IV Post Marketing Commitment Studies (PMCs), and as mentioned above if they are focused on safety they are also known as PASS. Phase IV PMCs are studies required of, or agreed to (i.e. 'negotiated'), by the sponsor at the time of NDA approval and this is particularly true of those drugs that have had accelerated approval. Phase IV clinical trials usually include a larger and more heterogeneous population than phase III trials with emphasis on reproducing usual clinical care conditions [24]. For some special populations, phase IV commitment trials represent a unique opportunity to determine the safety and efficacy of a drug [25]. This is particularly important for pediatric populations because only a small fraction of all drugs approved in the United States in the early 1990s have been studied in pediatric patients, and more than 70 % of new molecular entities were without pediatric labeling. Since 1994, the FDA began some initiatives to improve pediatric use information in drug labeling, by issuing a final rule revising the requirements for the Pediatric Use subsection of labeling (59 FR 64242, December 13, 1994). The regulation was designed to promote the inclusion of pediatric information from new clinical trials, as well as from previously published studies and case reports, in an effort to provide pediatric dosing and monitoring information in labeling; and, it required drug manufacturers to survey existing data and determine whether those data were sufficient to support additional pediatric use in a drugs labeling. On December 2nd, 1998 (23 FR 66632) the FDA issued the final rule (so-called pediatric rule) "Regulations Requiring Manufacturers to Assess the Safety and Effectiveness of New Drugs and Biological Products in Pediatric Patients". The pediatric rule was suspended by court order on October 17th 2002 [26]. Since then, many post-approval commitments have been required by the FDA and new clinical trials in pediatric populations are underway. It is evident that adequately designed phase IV clinical trials will impact drug utilization and prescriber's decisions particularly in children. For example, Lesko and Mitchell designed a practitioner-based, double blind, randomized trial in 27,065 children younger than 2 years old to compare the risk of serious adverse clinical events of ibuprofen versus acetaminophen suspension. They found an overall small risk of serious adverse events, but no difference by medication [27]. Phase IV commitments trials have also been used in exploratory special population studies, such as neonatal abstinence syndrome [28], and pregnant opiate-dependency [29, 30]. In those studies, the main research question focused on the efficacy and/or safety of a drug in small number of patients. For example, in the pregnant-opiate dependent study, Jones successfully transferred four drug-dependent pregnant inpatients from methadone to morphine and then buprenorphine [31].

An analysis of phase IV studies during 1987–1993 showed that each of the phase IV drugs had, on average, a commitment to conduct four studies [28]. The regulations regarding phase IV studies began in 1997 as part of the FDA Modernization Act. As a result of that act, the FDA was required to report annually on the status of postmarketing study commitments. In 1999 (a rule which became officially effective in 2001), the FDA published rules and formatting guidelines for the phase IV reports. Although these studies were a 'requirement' of NDA approval and are called 'commitment' studies, significant problems exist. In March 2006, the Federal Register reported on the status of postmarketing study commitments. Of 1,231 commitments, 787 were still pending (65 %), 231 were ongoing, and only 172 (14 %) were completed. The problem associated with these studies has been extensively discussed. For example, a recommendation by Public Citizen (a public advocacy group) followed the release of this FDA report, and noted that the FDA needs the ability to impose financial penalties as an incentive for drug companies to submit required annual post-market study reports on time. Peter Lurie, deputy director of Public Citizen's Health Research Group, told FDA news; '*The only thing the agency can do is take the drug off the market, which is a decision that often would not serve the public health very well*' [32]. In addition, the only mechanism that was available to remove a drug from the market was through a difficult legal channel. The FDA did not have the authority itself to withdraw a drug from the market, or suspend sales of a drug. In fact, the FDA could not even compel completion of a post-marketing study agreed upon at the time of approval, limit advertising of the drug, compel manufacturers to send out 'Dear Doctor' letters, or revise the product label of a drug without the approval of the company involved. Lurie noted that '*the great majority of postmarketing studies address safety issues, at least in part, so patients and physicians are denied critical safety information when these studies are not completed in a timely fashion.*' Lurie also criticized the FDA's report on the status of postmarketing commitments, noting there is no way of knowing what the deadlines are for each stage of the commitment and if they are being met or not, and for inadequate tracking systems for those who are initiating and those ongoing trials. In summary, in the past, the FDA set the schedule for firms to complete a battery of studies on products that require a phase IV study. The agency then evaluated each study to see if the drug company had fulfilled the requirements of the study commitment. If the company failed to submit data on time, the commitment was considered delayed. The reports were to contain information on the status of each FDA-required study specifically for clinical safety, clinical efficacy, clinical pharmacology, and non-clinical toxicology. The pharmaceutical firm then continued to submit the report until the FDA determined that the commitment had been fulfilled or that the agency no longer needed the reports.

In 2007, the FDA Amendments Act of 2007 was signed into law. Among other things, the Law addressed the need for ongoing evaluations of drug safety after drug approval, a way of addressing safety signals and performing high quality studies addressing those signals, new authority to require post marketing studies, civil penalties for non-compliance, the registration of all phase 2–4 trials, and the designation of some of the user's fees (10 %) to be earmarked for safety issues.

Practice Based Clinical Experience Studies

Physician Experience Studies (PES) were generally initiated by the pharmaceutical company that had marketed a particular drug. The name is descriptive of the intent of the study and in the past this was the type of study frequently associated with the term phase IV study. PES is generally not a RCT, and, therefore, has been most often criticized for its lack of scientific rigor. It does, however, in addition to providing physicians with experience in using a newly marketed drug, expose a large number of patients to that drug, potentially providing 'real world' information about the drugs adverse event profile.

An example of a PES is that of graded release diltiazem. The Antihypertensive Safety and Efficacy and Physician and Patient Satisfaction in Clinical Practice: Results from a Phase IV Practice-based Clinical Experience Trial with Diltiazem LA (DLA). The study enrolled a total of 139,965 patients with hypertension, and involved 15,155 physicians who were to perform a baseline evaluation and 2 follow-up visits [30]. Usual care treatment any other drug therapy was allowed as long as they were candidates for the addition of DLA. The potential to record efficacy and safety data for this large number of 'real world' patients was great. However, as a characteristic of these kinds of studies, only 50,836 (26 %) had data recorded for all three visits, and data on ADEs were missing for many as well. On the other hand, ADEs for 100,000 patients were available, and none of the ADEs attributed to DLA were reported in more than 1 % of patients, supporting the general safety profile of DLA. In recent years congressional oversight and increased criticism of this type of study with the accusation that it was merely a "marketing ploy" has all but stopped this type of approach.

Non FDA Studies

Non FDA mandated postmarketing studies may utilize the wide array of research designs, and should not be confused with PES studies. Examples of postmarketing studies include (1) RCTs with superiority testing, equivalence testing, large simple trials, 'phase V' trials and (2) surveillance studies such as effectiveness studies, drug utilization trials, epidemiologic observational studies (that concentrate on a safety profile of a drug), and classical RCTs. Not included in this present review is health services research, and health outcomes research that can also be studies of marketed drugs. Following is a discussion of some of the more common postmarketing research study designs. Postmarketing research falls under the umbrella of pharmacoepidemiologic studies (See Chap. 12).

Equivalence and Noninferiority Trials

As new drugs are finding it increasingly difficult to demonstrate superiority, equivalence trials are becoming more common. These trials are discussed in Chaps. 3 and 4.

Large Simple Trials

Not infrequently, an already marketed drug needs to be evaluated for a different condition than existed for its approval, or at a different dose, different release system, etc. In the aforementioned instance, the FDA might mandate a phase IV RCT that has all the characteristics of a classical phase III design. Some have suggested that this type of aforementioned study be termed a phase V study to distinguish it from the wide variety of other phase IV trials with all their attendant limitations and negative perceptions.

Table 5.4 RCTs: Rising complexity and burden of RCTs

Median/per protocol	2000–2003	2004–2007	2008–2011
Unique Procedures	20.5	28.2	30.4
Total Procedures	105.9	158.1	166.6
Work Burden (Units)	28.9	44.6	47.5
Eligibility Criteria	31	38	35
CRF Pages	55	180	169

National Academy of Sciences: Large Simple Trials; http://www.nap.edu/catalog.php?record_id=18400
RCT Randomized controlled trial, *CRF* Case report form

One type of postmarketing research is the Large Simple Trial (LST). The concept of large simple clinical trials has become more popular (as has been said the Large Simple Trial is an oxymoron no more). The idea is that it is increasingly necessary to just demonstrate modest benefits of an intervention, particularly in common conditions. The use of short-term studies, implemented in large populations is then attractive. In LSTs, the presumption is that the benefits are similar across participant types, so that the entry criteria can be broad, and the data entry and management can be simplified, and the cost thereby reduced (this then overcomes one of the limitations of the usual RCT –the homogeneity of the sample population). This model further depends on a relatively easily administered intervention and an easily ascertained outcome; but if these criteria are met, the size of the study also allows for a large enough sample size to assess less common ADE's.

The LST is also becoming more popular for certain phase III trials and RCTs in general. In 2013, the Institute of Medicine convened a panel that called for more LSTs. This report, acknowledged that one of the most persistent problems in medical care is the lack of evidence for clinical decisions, and one of the big issues has been the difficulty of answering the simplest questions (let alone solving complex ones). The Institute of Medicine report also defined LSTs broadly to encompass trials with simple randomization, broad eligibility, and enough participants to distinguish small to moderate effect; to focus on outcomes important to patient care, and to use simplified data collection. (The National Academies Press at http://www.nap.edu/catalog.php?record_id=18400) In large part the LST has also been suggested as a replacement for RCTs because of the increasing complexity and burden attendant with RCTs, a complexity that has been increasing over time (Table 5.4) An example of the organization for this type of trial is the Clinical Trial of Reviparin and Metabolic Modulation of Acute Myocardial Infarction (CREATE), as discussed by Yusuf et al. [33]. In this trial over 20,000 subjects from 21 countries were enrolled in order to compare two therapies-glucose-insulin-potassium infusion, and low molecular weight heparin.

Prospective, Randomized, Open-Label, Blinded Endpoint (PROBE) Design

A variation of the LST that also addresses a more 'real-world' scenario is the prospective randomized open-label blinded endpoint design (PROBE design). By using open-label therapy, the drug intervention and its comparator can be clinically titrated as would typically occur in a doctor's office (as compared to the fixed dosing regimens used in most RCTs). Of course, since it is open-label, blinding is lost with the PROBE design, but only as to the therapy. Blinding is maintained as to the ascertainment of the outcome. To test whether the use of open-label vs. double-blind therapy affected outcomes differentially, a meta-analysis of PROBE trials and double-blind trials in hypertension was reported by Smith et al. [34]. They found that changes in mean ambulatory blood pressure from double-blind controlled studies and PROBE trials were statistically equivalent.

Post-authorization (Surveillance) Studies

Pharmacovigilance deals with the detection, assessment, understanding and prevention of adverse effects or other drug-related problems. Traditionally, pharmacovigilance studies have been considered as part of the postmarketing phase of drug development because clinical trials in the premarketing phase are not powered to detect all adverse events, particularly uncommon adverse effects. It is known that in the occurrence of adverse drug reactions, other factors are involved such as the individual variation in pharmacogenetic profiles, drug metabolic pathways, the immune system, and drug-drug interactions. Additionally, the dose range established in clinical trials is not always representative of that used in the postmarketing phase. Cross et al. analyzed the new molecular entities approved by the FDA between 1980 and 1999 and they found that dosage changes occurred in 21 % of the approved entities, and of these, 79 % were related to safety. The median time to change following approval ranged from 1 to 15 years and the likelihood of a change in dosage was three times higher in new molecular entities approved in the 1990s compared to those approved in the 1980s [35] and, this would suggest that a wider variety of dosages and diverse populations need to be included in the premarketing phase and/or additional studies should be requested and enforced in the postmarketing phase. Further amplifying this point is a recent FDA news report [36] in which it was noted that there had been 45 Class I recalls (very serious potential to cause harm, injury, or death) in the last fiscal year (in many of the past years there had been only 1 or 2 such recalls) and also 193 class II recalls (potential to cause harm).

A clinical trial in 8,076 patients with rheumatoid arthritis that examined the association of rofecoxib (Vioxx) vs. naproxen on the incidence of gastrointestinal events, reported a higher percentage of incident myocardial infarction in the rofecoxib arm compared to naproxen during a median follow-up of 9 months that then questioned the drug safety of COX 2 inhibitors [37, 38]. The cardiac toxicity of rofecoxib was corroborated in a meta-analysis [39] database study [38], and in the APPROVe trial (Adenomatous Polyps Prevention on Vioxx) [40], a colorectal adenoma chemoprevention trial in which cardiovascular events were found to be associated with rofecoxib [38]. The APPROVe trial is an example of phase IV trial that was organized for another potential indication of rofecoxib, the reduction of the risk of recurrent adenomatous polyps among patients with a history of colorectal adenomas. In that multicenter, randomized, placebo-controlled, double-blind study, 2,600 patients with history of colorectal adenoma was enrolled, but after 3,059 patient-years of follow-up there was an increased risk of cardiovascular events. All of the above evidence resulted in the final decision of the manufacturer to withdraw rofecoxib from the market [41].

The type of scandals that are associated with drug safety and the pressure of society, have contributed to the development of initiatives for performing more pharmacovigilance studies. Some countries, for example, are now requiring manufacturers to monitor the adverse drug events of approved medications. In France, manufacturers must present a pre-reimbursement evaluation and a postmarketing impact study [42]. In fact, France has a policy for the overall assessment of the public health impact of new drugs [42].

In the United States, the recent withdrawals from the market (particularly for drugs that were approved through the expedited process by the FDA) indicate a need to start pharmacovigilance programs at the earliest stages of drug development, encouraging the identification of safety signals, risk assessment, and communication of those risks. The FDA has started developing algorithms to facilitate detection of adverse-event signals using the 'MedWatch', a spontaneous reporting adverse event system, to institute risk-management measures.

The 'MedWatch' is a voluntary system where providers, patients or manufacturers can report serious, undesirable experiences associated with the use of a medical product in a patient. An event is considered serious if it is associated with patient's death or increases the risk of death; the patient requires hospitalization, the product causes disability, a congenital anomaly occurs, or the adverse event requires medical or surgical intervention to prevent permanent impairment or damage [15]. The main obstacle of MedWatch is the high rate of underreporting adverse drug reactions which is then translated into delays in detecting adverse drug reactions [43, 44]. Adverse events that are associated with vaccines or with veterinary products are not reported to the Medwatch. The FDA collates the Medwatch reports and determines if more research is needed to establish a cause-effect relationship between the drug and the adverse event. Then, the FDA defines the actions that manufacturers, providers, and patients should take.

Another consequence from the recent drug withdrawals is the release of more safety information from the FDA to the public and press, as well as the creation of a new board to help monitor drugs [45]. In 2012, there were 65 MEDWATCH Drug Safety Alerts (e.g. the FDA notified healthcare professionals of possible risks when using blood pressure medicines containing aliskiren with other drugs called angiotensin converting enzyme inhibitors (ACEIs) and angiotensin receptor blockers (ARBs) in patients with diabetes or kidney (renal) impairment); and, 47 device alerts (e.g. four reports of incidents, one resulting in patient injury, in which a device apparently malfunctioned resulting in difficulty deflating a balloon after the surgeon pulled against resistance in response to the balloon moving distally during dilation.

The force applied to the catheter stretched and narrowed the catheter shaft, causing the balloon to be difficult or impossible to deflate).

Finally, the FDA has established a "safety first initiative". The Safety First Initiative ensures drug safety throughout the drug lifecycle by giving pre-marketing drug review and post-marketing safety equal focus. This calls for inter-office, multidisciplinary safety-issue teams to assess significant safety issues, recommend actions, and monitor sponsors' activities.

Effectiveness (Pragmatic) Clinical Trials

As mentioned before, one of the limitations of phase 3 RCTs is their limited generalizability. Although the RCT may be the best way to evaluate efficacy under optimal conditions, it may not accurately reflect the drugs effectiveness under usual case ('real world') conditions. That is many phase 3 RCTs ignore existing treatments as the comparator (i.e. use a placebo control) and do not examine if a new treatment is better than the existing approved treatment(s), they key question for clinical practice. Clearly, clinical practice would follow evidence-based medicine, which is derived from the RCT and meta-analyses of RCTs. But often the outcomes of clinical practice are not equal to that of the RCTs (due to differences in patients, the quality of the other treatments they receive, drug-drug and drug-disease interactions they may experience-these being much more common in the heterogeneity of clinical practice patients compared to the highly selected clinical trial patients). In addition, phase 3 trials are not only interested in efficacy but also biologic mechanisms for how the treatment works and often measure intermediate or surrogate outcomes that may be less relevant to patient care. It is in this aforementioned setting that Effectiveness (Pragmatic) Trials are increasingly important. The term "pragmatic clinical trial" is credited to Schwartz and Lellouch [46]. They made the point that when comparisons are made between two treatment groups, the problem is often inadequately specified in its overall characteristics and thus may represent at least one of two approaches: the first corresponding to an explanatory approach aimed primarily at "understanding"; and, the second aimed at "decision making". That is, the explanatory approach seeks to discover whether a difference exists between two treatments (one usually being a comparator placebo) that are specified by strict and usually simple definitions, while the second (the pragmatic approach) aims to answer the question "which of the two treatments would be preferred where the treatments are complex and flexible". As such, explanatory trials often enroll homogeneous patients with few comorbid conditions in an attempt to reduce response variation, while pragmatic trials have fewer inclusion/exclusion criteria and thus enroll a more heterogeneous population.

As is true of any decision made in research, there are always trade-offs (compromises) one has to make. While effectiveness/pragmatic trials may increase generalizability, it does so at the expense of internal validity. Also, because pragmatic trials

Table 5.5 Comparison of efficacy/explanatory vs. effectiveness/pragmatic clinical trials

	Efficacy/explanatory	Effectiveness/pragmatic
Objective	Optimal efficacy	Usual effectiveness
Motivation	FDA approval	Formulary approval
Intervention	Fixed regimen	Flexible regimen
Comparator	Placebo	Currently used treatments
Design	RCT	Open-label
Subjects	Highly selected, compliant	"All comers"
Outcomes	Condition of interest	Comprehensive
Duration	Short-term, surrogate endpoint	Long-term
Other	Mechanism of action	Real world performance

enroll a heterogeneous population, they frequently require larger sample sizes. Table 5.5 contrasts important considerations between efficacy and effectiveness studies. Additional differences include: explanatory trial interventions are usually delivered by highly trained, specialized and skilled practitioners, while pragmatic trials mimic routine practice and the practitioner in these trials are skilled in routine clinical practice, but may not be as skilled in clinical research. An example of some of these issues was reported by Taylor et al. [47]. The British Association for Cardiac Rehabilitation performs an annual questionnaire of the 325 cardiac rehabilitation programs in the UK. Taylor et al. compared the patient characteristics and program details of this survey with RCTs included in the 2004 Cochrane review. They found 'considerable differences' between the RCTs of cardiac rehabilitation and the actual practice in the UK (Table 5.6), differences suggesting that the real world practice of cardiac rehabilitation is unlikely to be as effective as clinical trials would suggest.

Thorpe et al. have suggested an approach to evaluate how explanatory or pragmatic a particular study is (since many studies are a mixture of the two) by developing a pragmatic-explanatory continuum indicator summary (PRECIS) based upon ten domains that differentiate the two to include: practitioner expertise, follow-up intensity, outcomes, participant compliance, flexibility of the comparison intervention, flexibility of the experimental intervention, eligibility criteria, and primary analysis. For each "spoke" a judgment is made on a continuum and the dots connected, providing a visual of how pragmatic a trial is (Table 5.7). They do note that this is a first step in identifying the degree to which these trials are pragmatic [48].

Drug Utilization and Pharmacoeconomic Studies

One of the main reasons to conduct postmarketing studies is to demonstrate the economic efficiency of prescribing a new drug. In this instance, the manufacturer is interested in showing the relationship of risks, benefits and costs involved in the use

Table 5.6 A comparison from the UK of differences between RCTs of cardiac rehabilitation and actual practice

	Cochrane report	Rehabilitation survey	Prevention survey
Population characteristics			
Mean age	54	64	Unknown
% Women	10.4	26.4	Unknown
% MI	86	54	Unknown
% CABG	6	24	
% PTCA	5	13	
Intervention characteristics			
% Exercise only	39	0	0
Duration (wks)	18	7.5	7
Exercise duration	58	Unknown	60
Frequency/wk	2.8	1.66	1.67
% VO2	75	Unknown	Unknown
# of sessions	50	12.4	12
% Hospital based	91	66	100

Adapted from: Taylor et al. [47]

MI myocardial infarction, *CABG* coronary artery bypass grafting, *PTCA* percutaneous transluminal coronary artery angioplasty

Table 5.7 Factors that help determine if a study meets a pragmatic design

The pragmatic-explanatory continuum
Practitioner expertise in experimental research
Flexibility of the experimental intervention
Eligibility criteria
Primary analysis
Practitioner adherence
Participant compliance
Outcomes
Follow up intensity
Practitioner expertise with the comparison group
Flexibility of the comparison intervention

Adapted from: Thorpe et al. [48]

of a new drug in order to show the value for the products cost. That value is essential for decision makers and prescriber's, who will select medications for formularies or prescribe the most appropriate medication for patients.

Most of the pharmacoeconomic studies have been carried out in the postmarketing phase using modeling techniques. Simulation models are mathematical abstractions of reality, based on both assumptions and judgments [49]. Those models are built using decision analysis, state transition modeling, discrete event simulation, and survival modeling techniques [50]. The aforementioned models could allow for the adjustment of various parameters in outcomes and costs, and could explore the effect of changes in healthcare systems and policies if they clearly present and validate the assumptions made. Unfortunately, many economic models have issues related to model building, model assumptions, and lack of data, that limits their acceptability by decision makers and consumers.

One of the issues for simulated models is that they usually get information from different sources. For example, a cost-effectiveness model of antidiabetic medications obtained information from the literature and expert panels to determine algorithms of treatment; success, failures and adverse events were obtained from product labeling, literature and the drugs NDA; resource utilization data (i.e. physician office visits, laboratory tests, eye exams, etc.) were acquired from the American Diabetes Association guidelines, and costs were obtained from the literature [51]. This mixture of heterogeneous information raises questions related to the validity of the model. As a potential solution, some manufacturers have started including pharmacoeconomic evaluations alongside clinical trials. This 'solution' might appear logical but the approach has limitations, such as the difficulty in merging clinical and economic outcomes in one study, limitations regarding the length of the trial as these may differ for the clinical vs. the economic measures, differing sample size considerations, and finally differences in efficacy vs. effectiveness.

Frequently, trials are organized to show the efficacy of new medications but most phase II (and for that matter phase III) trials use surrogate measures as their primary end points and the long-term efficacy of the drug is unknown. For example, glycosylated hemoglobin (HbA_{1c}) or fasting plasma glucose is frequently used as an indicator of drug efficacy for phase II or phase III trials. However, when those efficacy data are used for simulation models, there is a lack of long-term efficacy information that then requires a series of controversial assumptions. To overcome that latter issue, economists are focusing on short- term models adducing that health maintenance organizations (HMOs) are more interested in those outcomes while society is interested in both short and long-term outcomes. For example, a decision-tree model was developed to assess the direct medical costs and effectiveness of achieving glycosylated hemoglobin (HBA_{1c}) values with antidiabetic medications during the first 3 years of treatment. The authors justified the short-term period arguing that it was more relevant for decision makers to make guideline and formulary decisions [51]. Although it may look easy to switch short-term for long-term outcomes this switch may be problematic, because short-term outcomes may not reflect long term outcomes. Another factor to consider is the length of a trial because if there is a considerable lag-time between premarketing trials and postmarketing trials, practice patterns may have changed thereby affecting HMO decisions.

The size of the trial is also a very important factor to take into account in pharmacoeconomics because trials are powered for clinical outcomes and not for economic outcomes. If economic outcomes are used to power a trial, then a larger sample size will be required because economic outcomes have higher variation than clinical outcomes [52].

In addition, the use of surrogate outcomes may not be economically relevant, a factor that needs to be considered by health economists and trialists during a trials planning phase. A question could then arise: could costs be used as endpoints? The short-answer is no, because costs data are not sensitive surrogates endpoints since cost and clinical outcomes may be disparate.

Finally, the efficiency of a new product requires that the manufacturer demonstrate the effectiveness of the product, that is, how the product behaves in the real world and not under 'experimental' conditions (efficacy). For example, the manufacturers may want to show that the new product is more cost-effective than current therapies or at least as good as new alternatives, but they need real-life data which are almost always absent when that product is launched. This is an important issue because premarketing trials are usually carried out in selective sites that are not representative of the practice community at large. Why are 'real' data so important? It is known that once a product is in the market, there is a wide variation in how the product is used by providers (e.g. indications, target population – different age, gender, socioeconomic status, patients with co-morbidities or multiple medications; adherence to medical guidelines, and variation among providers), or used by patients (e.g. patient adherence to medications, variation in the disease knowledge, access to care, and type of care). Additionally, the new product might prompt changes in the resource utilization for a particular disease. For example, when repaglinide was introduced into the market, it was recommended that in patients with type 2 diabetes postprandial and fasting glucose, as well as HbA_{1c} be monitored [53, 54], this type of monitoring would require testing that is additional to the usual management of patients with diabetes. As someone once said, *"use the new drugs now, while they still work."*

Comparative Effectiveness Research (CER)

Comparative effectiveness research is defined by the Institute of Medicine committee as *"the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care"*. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels. Comparative effectiveness usually compares two or more types of treatment, such as different drugs, for the same disease. Comparative effectiveness also can compare types of surgery or other kinds of medical procedures and tests.

Three commonly used approaches used to look for "safety signals" after a drug has been approved for the market are:

- Meta-analysis of clinical trials (See Chap. 10)
- Disproportionality methods using the FDAs (or other data bases) adverse event reporting systems (AERS)
- Observational analyses of administrative data bases (i.e. tapping into large health care data bases)

The comparison of drugs or devices can generated from already performed research studies (i.e. systematic reviews), or by conducting studies that generate new evidence of effectiveness or comparative effectiveness of a test, treatment,

Table 5.8 Seven steps that are involved in conducting comparative effectiveness research

Identify new and emerging clinical interventions
Review and synthesize current medical research
Identify gaps between existing medical research and the needs of clinical practice
Promote and generate new scientific evidence and analytic tools
Train and develop clinical researchers
Translate and disseminate research findings to diverse stakeholders
Reach out to stakeholders via a citizens forum

procedure, or health-care service. The pragmatic (effectiveness) trial is the "bread and butter" design of CER (see above).

The Agency for Healthcare Research and Quality (AHRQ) have suggested seven steps that are involved in conducting CER as follows (Table 5.8):

- Identify new and emerging clinical interventions.
- Review and synthesize current medical research.
- Identify gaps between existing medical research and the needs of clinical practice.
- Promote and generate new scientific evidence and analytic tools.
- Train and develop clinical researchers.
- Translate and disseminate research findings to diverse stakeholders.
- Reach out to stakeholders via a citizen's forum [55].

Statistical findings show that there is large gap in the quality and outcomes and health services being delivered with a significant geographic variation such that patients in the highest-spending regions of the country receive 60 % more health services than those in the lowest-spending regions, yet this additional care is not in general, associated with improved outcomes. Some of the metrics used in CER are cost effective analysis; the incremental cost-effectiveness ratio (ICER) given by the difference in costs between two health programs divided by the difference in outcomes between the programs; and, the quality-adjusted life year (QALY), a measure of disease burden, including both the quality and the quantity of life lived as a unit for measuring the health gain of an intervention calculated as the number of years of life saved and adjusted for quality generally applied to the denominator in the ICER.

There are at least two controversies over the use of CER to determine treatment:

- There are some questions among professionals about whether cost should be one of the data points studied in CER. For example, suppose Treatment A turns out to be much more effective than Treatment B, but costs twice as much. If you include the cost of the treatment in the consideration of its comparison, then Treatment B might be considered to be more effective. Should that be the conclusion? There are no easy answers to this question.

- Some patients and professionals are concerned that CER results would then eliminate some possibilities for patients, which might be considered a form of rationing.

Although CER is a burgeoning field, further discussion is beyond the scope of this book and the interested reader can refer to many excellent reviews of this area [56].

Because of the aforementioned issues, economic data alongside (or merged with) clinical trials are important because data obtained in premarketing trials could shed light on the goal of anticipating results in postmarketing trials, they could contribute to developing cost weights for future studies, and they could help to identify the resources that have the highest impact of the new drug.

Discussion

The term ‘phase IV study’ has become misunderstood and has taken on negative connotations that have led some experts to question the validity of such trials. Pocock emphasizes this latter point – ‘*such a trial has virtually no scientific merit and is used as a vehicle to get the drug started in routine medical practice*’ [57]. He was undoubtedly referring to phase IV physician experience studies at the time. But the phase IV PES may have had some merit, even given that adverse event reporting is voluntary, and underreporting of events is believed to be common (this is contrast to phase III trials where UAEs are arguably over reported). It is true that many phase IV studies have limitations in their research design, that the follow-up of patients enrolled in phase IV trials may be less rigorous than in controlled clinical trials (which can decrease the quantity and quality of information about the safety and efficacy of the medication being evaluated) [58, 59] but, due to the highly varied designs of phase IV studies, the utility of the information they provide will vary substantially from one study to another.

Due to the limitations of the current system for identifying adverse events, Strom has suggested a paradigm shift from the current traditional model of drug development and approval. He supports this paradigm shift based upon the fact that ‘…51 % of drugs have label changes because of safety issues discovered after marketing, 20 % of drugs get a new black box warning after marketing, and 3–4 % of drugs are ultimately withdrawn for safety reasons.’ The FDA website lists 12 drugs withdrawn from the market between 1997 and 2001 as shown in Table 5.9.

Strom’s suggested paradigm for studying drug safety has a shortened phase III program followed by conditional approval during which time, required postmarketing studies would need to be performed (and the FDA would need to be given the power to regulate this phase in the same manner that they now have with phase I-III studies). He further recommends that once the conditional approval phase has ascertained safety in an additional 30,000 or more patients, the current system of optional and/or unregulated studies could be performed (Table 5.10).

Table 5.9 The current (traditional) vs. some proposed paradigms for drug development

Approximate number of participants per phase					
Models	Phase 1	Phase 2	Phase 3	Phase 4A	Phase 4B
Traditional	0–100	100–500	500–2,000	>3,000	NA
Evolving	0–100	100–500	500–10,000	>10,000	NA
Proposed	1–100	100–500	500–3,000	20,000–300,000	>300,000

Phase 4A refers to conditional approval, 4B to full approval

The "conditional approval concept" has been supported by the Institute of Medicine and goes even further. The Institute of Medicine proposes to include a symbol for new drugs, new combinations of active substances, and new systems of delivery of existing drugs in the product label. This symbol would last 2 years and it would indicate the conditional approval of a drug until enough information from postmarketing surveillance is available, and during this period, the manufacturer would limit the use of direct-to-consumer advertising [60]. The question is how much impact this would have on prescriber's since some studies have shown that prescriber's often fail to follow black box warnings labels [61]. The Institute of Medicine also recommends that the FDA should reevaluate cumulative data on safety and efficacy no later than 5 years after approval. However, these changes are expected to have low impact if they are not accompanied by changes in law. Currently, the FDA has authority under the Food and Drug Administration Modernization Act of 1997 to require sponsors to submit annual updates and progress of study commitments. However, the FDA did not have the authority and tools to enforce those commitments until 2006 when Congress gave special authority to enforce these commitments.

It is also important not to lump the phase IV study with other postmarketing research, research that may be every bit as scientifically rigorous as that associated with RCTs. Postmarketing studies are essential to establish patterns of physician prescribing and patient drug utilization and they are usually carried out using observational designs. Investigators frequently relate postmarketing surveillance studies with pharmacovigilance studies, and this might be a signal of what is happening in practice. In the last 25 years, 10 % of the new drugs marketed in the United States have been withdrawn or were the subject of major warnings about serious or life-threatening side effects during the postmarketing phase. This situation has called for concrete actions such as closer monitoring of new drugs, the development of better notification systems for adverse events, and presentation of transparent and high quality data.

Clinical pharmacologists and pharmacoepidemiologists are trying to promote the collection of blood samples at the population level for pharmacokinetic analysis. A study in psychiatric inpatients treated with alprazolam collected two blood samples at different time intervals to assess the pharmacokinetic variability of heterogeneous patient populations [62]. This information could contribute to establishing dosages and frequency of drug administration in patients with co-morbidities,

Table 5.10 Drugs withdrawn from the market between 1950 and 2011. Significant withdrawals

Drug name	Withdrawn	Remarks
Thalidomide	1950s–1960s	Withdrawn because of risk of teratogenicity; returned to market for use in leprosy and multiple myeloma under FDA orphan drug rules
Lysergic acid diethylamide (LSD)	1950s–1960s	Marketed as a psychiatric drug; withdrawn after it became widely used recreationally
Diethylstilbestrol	1970s	Withdrawn because of risk of teratogenicity
Phenformin and buformin	1978	Withdrawn because of risk of lactic acidosis
Ticrynafen	1982	Withdrawn because of risk of hepatitis
Zimelidine	1983	Withdrawn worldwide because of risk of Guillain-Barré syndrome
Phenacetin	1983	An ingredient in "A.P.C." tablet; withdrawn because of risk of cancer and kidney disease
Methaqualone	1984	Withdrawn because of risk of addiction and overdose
Nomifensine (Merital)	1986	Withdrawn because of risk of hemolytic anemia
Triazolam	1991	Withdrawn in the United Kingdom because of risk of psychiatric adverse drug reactions. This drug continues to be available in the U.S.
Terodiline (Micturin)	1991	Prolonged QT interval
Temafloxacin	1992	Withdrawn in the United States because of allergic reactions and cases of hemolytic anemia, leading to three patient deaths.
Flosequinan (Manoplax)	1993	Withdrawn in the United States because of an increased risk of hospitalization or death
Alpidem (Ananxyl)	1995	Not approved in the US, withdrawn in France in 1994 and the rest of the market in 1995 because of rare but serious hepatotoxicity
Chlormezanone (Trancopal)	1996	Withdrawn because of rare but serious cases of toxic epidermal necrolysis
Fen-phen (popular combination of fenfluramine and phentermine)	1997	Phentermine remains on the market, dexfenfluramine and fenfluramine – later withdrawn as caused heart valve disorder
Tolrestat (Alredase)	1997	Withdrawn because of risk of severe hepatotoxicity
Terfenadine (Seldane, Triludan)	1998	Withdrawn because of risk of cardiac arrhythmias; superseded by fexofenadine
Mibefradil (Posicor)	1998	Withdrawn because of dangerous interactions with other drugs
Etretinate	1990s	Risk of birth defects; narrow therapeutic index
Tolcapone (Tasmar)	1998	Hepatotoxicity
Temazepam (Restoril, Euhypnos, Normison, Remestan, Tenox, Norkotral)	1999	Withdrawn in Sweden and Norway because of diversion, abuse, and a relatively high rate of overdose deaths in comparison to other drugs of its group. This drug continues to be available in most of the world including the U.S., but under strict controls.

(continued)

Table 5.10 (continued)

Drug name	Withdrawn	Remarks
Astemizole (Hismanal)	1999	Arrhythmias because of interactions with other drugs
Grepafloxacin (Raxar)	1999	Prolonged QT interval
Levamisole (Ergamisol)	1999	Still used as veterinary drug; in humans was used to treat melanoma before it was withdrawn for agranulocytosis
Troglitazone (Rezulin)	2000	Withdrawn because of risk of hepatotoxicity; superseded by pioglitazone and rosiglitazone
Alosetron (Lotronex)	2000	Withdrawn because of risk of fatal complications of constipation; reintroduced 2002 on a restricted basis
Cisapride (Propulsid)	2000s	Withdrawn in many countries because of risk of cardiac arrhythmias
Amineptine (Survector)	2000	Withdrawn because of hepatotoxicity, dermatological side effects, and abuse potential
Phenylpropanolamine (Propagest, Dexatrim)	2000	Withdrawn because of risk of stroke in women under 50 years of age when taken at high doses (75 mg twice daily) for weight loss
Trovafloxacin (Trovan)	2001	Withdrawn because of risk of liver failure
Cerivastatin (Baycol, Lipobay)	2001	Withdrawn because of risk of rhabdomyolysis
Rapacuronium (Raplon)	2001	Withdrawn in many countries because of risk of fatal bronchospasm
Nefazodone	2003	Branded version withdrawn by originator in several countries in 2003, and in the US and Canada in 2004 for hepatotoxicity. Generic versions available
Rofecoxib (Vioxx)	2004	Withdrawn because of risk of myocardial infarction
Co-proxamol (Distalgesic)	2004	Withdrawn in the UK due to overdose dangers
mixed amphetamine salts (Adderall XR)	2005	Withdrawn in Canada because of risk of stroke. See Health Canada press release. The ban was later lifted because the death rate among those taking Adderall XR was determined to be no greater than those not taking Adderall
hydromorphone extended-release (Palladone)	2005	Withdrawn because of a high risk of accidental overdose when administered with alcohol
Valdecoxib (Bextra)	2005	Withdrawn in US due to concerns about heart attack and stroke
Thioridazine (Melleril)	2005	Withdrawn from U.K. market because of cardiotoxicity
Pemoline (Cylert)	2005	Withdrawn from U.S. market because of hepatotoxicity

(continued)

Table 5.10 (continued)

Drug name	Withdrawn	Remarks
Natalizumab (Tysabri)	2005–2006	Voluntarily withdrawn from U.S. market because of risk of Progressive multifocal leukoencephalopathy (PML). Returned to market July, 2006
Ximelagatran (Exanta)	2006	Withdrawn because of risk of hepatotoxicity (liver damage)
Pergolide (Permax)	2007	Voluntarily withdrawn in the U.S. because of the risk of heart valve damage. Still available elsewhere
Tegaserod (Zelnorm)	2007	Withdrawn because of imbalance of cardiovascular ischemic events, including heart attack and stroke. Was available through a restricted access program until April 2008
Aprotinin (Trasylol)	2007	Withdrawn because of increased risk of complications or death; permanently withdrawn in 2008 except for research use
Inhaled insulin (Exubera)	2007	Withdrawn in the UK due to poor sales caused by national restrictions on prescribing, doubts over long term safety and too high a cost
Lumiracoxib (Prexige)	2007–2008	Progressively withdrawn around the world because of serious side effects, mainly liver damage
Rimonabant (Acomplia)	2008	Withdrawn around the world because of risk of severe depression and suicide
Efalizumab (Raptiva)	2009	Withdrawn because of increased risk of progressive multifocal leukoencephalopathy
Sibutramine (Reductil/Meridia)	2010	Withdrawn in Europe, Australasia, Canada, and the U.S. because of increased cardiovascular risk
Gemtuzumab ozogamicin (Mylota)	2010	Withdrawn in the U.S. due to increased risks of veno-occlusive disease and based on results of a clinical trial in which it showed no benefit in acute myeloid leukemia (AML)
Propoxyphene (Darvocet/Darvon)	2010	Withdrawn from worldwide market because of increased risk of heart attacks and stroke
Rosiglitazone (Avandia)	2010	Withdrawn in Europe because of increased risk of heart attacks and death. This drug continues to be available in the U.S.
Drotrecogin alfa (Xigris)	2011	Withdrawn by Lily worldwide following results of a study that showed lack of efficacy

From Wikipedia, the free encyclopedia www.wikipedia.com. Accessed 23 Apr 2013 those treated with multiple medications and special populations. Clearly, the rubric of the phase IV study has taken on an expanded and meaningful role in drug development, use, and safety.

Appendix: The Following Definitions Were Used in This Manuscript

Definitions of phase IV trials:

- Post-marketing studies to delineate additional information including the drug's risks, benefits, and optimal use. clinicaltrials.mayo.edu/glossary.cfm
- Postmarketing studies, carried out after licensure of the drug. Generally, a phase IV trial is a randomized, controlled trial that is designed to evaluate the long-term safety and efficacy of a drug for a given indication. Phase IV trials are important in evaluating AIDS drugs because many drugs for HIV infection have been given accelerated approval with small amounts of clinical data about the drugs' effectiveness. www.amfar.org/cgi-bin/iowa/bridge.html
- In medicine, a clinical trial (synonyms: clinical studies, research protocols, medical research) is a research study. en.wikipedia.org/wiki/Phase_IV_trials

1. adverse drug event or adverse drug experience: 'an untoward outcome that occurs during or following clinical use of a drug, whether preventable or not' (does not mention causality)
2. adverse experience: 'any adverse event associated with the use of a drug or biological product in humans, whether or not considered product related' (causality not assumed)
3. adverse drug reaction: 'an adverse drug event that is judged to be caused by the drug' (specifically refers to causality)
4. 'Studies of adverse effects examine case reports of adverse drug reactions, attempting to judge subjectively whether the adverse events were indeed caused by the antecedent drug exposure' (specifically focuses on causality)
5. 'Studies of adverse events explore any medical events experienced by patients and use epidemiologic methods to investigate whether any given event occurs more often in those who receive a drug than in those who do not receive the drug' (a bit equivocal about causality: positive association v. causal association).

'Pharmacovigilance is a type of continual monitoring for unwanted effects and other safety-related aspects of drugs that are already on the market. In practice, pharmacovigilance refers to the spontaneous reporting systems that allow health care professionals and others to report adverse drug reactions to a central agency. The central agency can then combine reports from many sources to produce a more informative safety profile for the drug product than could be done based on one or a few reports from one or a few health care professionals.'

References

1. Glasser SP, Salas M, Delzell E. Importance and challenges of studying marketed drugs: what is a phase IV study? Common clinical research designs, registries, and self-reporting systems. J Clin Pharmacol. 2007;47:1074–86.
2. Brainy Quote. Wernher von Braun Quotes. http://www.brainyquote.com/quotes/authors/w/wernher_von_braun.html. Accessed 15 July 2013.
3. Hartzema A. Pharmacoepidemiology. 3rd ed. Cincinnati: Harvey Whitney Books Company; 1998.
4. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ. 1997;315:1059.
5. Corrigan OP. A risky business: the detection of adverse drug reactions in clinical trials and post-marketing exercises. Soc Sci Med. 2002;55:497–507.
6. Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. JAMA. 1992;268:1417–22.
7. Simon SD. Is the randomized clinical trial the gold standard of research? J Androl. 2001;22:938–43.
8. Farahani P, Levine M, Gaebel K, Thabane L. Clinical data gap between phase III clinical trials (pre-marketing) and phase IV (post-marketing) studies: evaluation of etanercept in rheumatoid arthritis. Can J Clin Pharmacol. 2005;12:e254–63.
9. Gough S. Post-marketing surveillance: a UK/European perspective. Curr Med Res Opin. 2005;21:565–70.
10. Gex-Fabry M, Balant-Gorgia AE, Balant LP. Therapeutic drug monitoring databases for post-marketing surveillance of drug-drug interactions. Drug Saf. 2001;24:947–59.
11. Kaufman DW, Kelly JP, Rosenberg L, Anderson TE, Mitchell AA. Recent patterns of medication use in the ambulatory adult population of the United States: the Slone survey. JAMA. 2002;287:337–44.
12. Vijan S, Kent DM, Hayward RA. Are randomized controlled trials sufficient evidence to guide clinical practice in type II (non-insulin-dependent) diabetes mellitus? Diabetologia. 2000;43:125–30.
13. Castle WM, Lewis JA. Postmarketing surveillance of adverse drug reactions. BMJ. 1984;288:1458–9.
14. Hayward RA, Kent DM, Vijan S, Hofer TP. Reporting clinical trial results to inform providers, payers, and consumers. Health Aff (Millwood). 2005;24:1571–81.
15. MedWatch: Voluntary Reporting by Health Professionals. http://www.fda.gov/medwatch/report/hcp.htm. Accessed 12 Oct 2006.
16. European Parliament and the Council of the European Union. Directive 2010/84/EU of the European Parliament of the Council of 15. Off J Eur Union. Strasbourg, Germany; 2010.
17. Giezen TJ, Mantel-Teeuwisse AK, Straus SM, Egberts TC, Blackburn S, Persson I, et al. Evaluation of post-authorization safety studies in the first cohort of EU Risk Management Plans at time of regulatory approval. Drug Saf. 2009;32:1175–87. doi:10.2165/11318980-000000000-00000.
18. Edwards C, Blowers DA, Pover GM. Fosinopril national survey: a post-marketing surveillance study of fosinopril (Staril) in general practice in the UK. Int J Clin Pract. 1997;51:394–8.
19. Fallowfield JM, Blenkinsopp J, Raza A, Fowkes AG, Higgins TJ, Bridgman KM. Post-marketing surveillance of lisinopril in general practice in the UK. Br J Clin Pract. 1993;47:296–304.
20. Marsh BT, Atkins MJ, Talbot DJ, Fairey IT. A post-marketing acceptability study in 11,685 patients of the efficacy of timolol/bendrofluazide in the management of hypertension in general practice. J Int Med Res. 1987;15:106–14.
21. Riley J, Wilton LV, Shakir SA. A post-marketing observational study to assess the safety of mibefradil in the community in England. Int J Clin Pharmacol Ther. 2002;40:241–8.
22. Schmidt J, Kraul H. Clinical experience with spirapril in human hypertension. J Cardiovasc Pharmacol. 1999;34 Suppl 1:S25–30.

23. Ueng KC, Chen ZC, Yeh PS, Hung KC, Hu SA, Hung YJ, et al. Nifedipine OROS in Chinese patients with hypertension – results of a post-marketing surveillance study in Taiwan. Blood Press Suppl. 2005;1:32–8.
24. Tognoni G, Alli C, Avanzini F, Bettelli G, Colombo F, Corso R, et al. Randomised clinical trials in general practice: lessons from a failure. BMJ. 1991;303:969–71.
25. Ben-Menachem E. Data from regulatory studies: what do they tell? What don't they tell? Acta Neurol Scand Suppl. 2005;181:21–5.
26. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (DCER), Center for Biologics Evaluation and Research (CBER). Guidance for industry and review staff: pediatric information incorporated into human prescription drug and biological products labeling, good review practice. In: U.S. Department of Health & Human Services, editor. Rockville: Office of Communication, Outreach, and Development, HFM-40, Center for Biologics Evaluation and Research, Food and Drug Administration; 2013.
27. Lesko SM, Mitchell AA. The safety of acetaminophen and ibuprofen among children younger than two years old. Pediatrics. 1999;104:e39.
28. Jackson L, Ting A, McKay S, Galea P, Skeoch C. A randomised controlled trial of morphine versus phenobarbitone for neonatal abstinence syndrome. Arch Dis Child Fetal Neonatal Ed. 2004;89:F300–4.
29. Fischer G, Ortner R, Rohrmeister K, Jagsch R, Baewert A, Langer M, et al. Methadone versus buprenorphine in pregnant addicts: a double-blind, double-dummy comparison study. Addiction. 2006;101:275–81.
30. Vocci F, Ling W. Medications development: successes and challenges. Pharmacol Ther. 2005;108:94–108.
31. Jones HE, Suess P, Jasinski DR, Johnson RE. Transferring methadone-stabilized pregnant patients to buprenorphine using an immediate release morphine transition: an open-label exploratory study. Am J Addict. 2006;15:61–70.
32. FDA report highlights poor enforcement of post-marketing follow-up. http://www.citizen.org/pressroom/release.cfm?ID=2147. Accessed 12 Oct 2006.
33. Yusuf S, Mehta SR, Xie C, Ahmed RJ, Xavier D, Pais P, et al. Effects of reviparin, a low-molecular-weight heparin, on mortality, reinfarction, and strokes in patients with acute myocardial infarction presenting with ST-segment elevation. JAMA. 2005;293:427–35.
34. Smith DH, Neutel JM, Lacourciere Y, Kempthorne-Rawson J. Prospective, randomized, open-label, blinded-endpoint (PROBE) designed trials yield the same results as double-blind, placebo-controlled trials with respect to ABPM measurements. J Hypertens. 2003;21:1291–8.
35. Cross J, Lee H, Westelinck A, Nelson J, Grudzinskas C, Peck C. Postmarketing drug dosage changes of 499 FDA-approved new molecular entities, 1980–1999. Pharmacoepidemiol Drug Saf. 2002;11:439–46.
36. FDA News. Drug Daily Bull. 2006;3, No 207.
37. Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. N Engl J Med. 2000;343:1520–8, 2 p following 8.
38. Mukherjee D, Nissen SE, Topol EJ. Risk of cardiovascular events associated with selective COX-2 inhibitors. JAMA. 2001;286:954–9.
39. Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet. 2004;364:2021–9.
40. Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med. 2005;352:1092–102.
41. Merck and Company. http://www.vioxx.com/vioxx/documents/english/vioxx_press_release.pdf. Accessed 4 Oct 2006.
42. Abenhaim L. Lessons from the withdrawal of rofecoxib: France has policy for overall assessment of public health impact of new drugs. BMJ. 2004;329:1342.

43. van Grootheest A, de Graafe L, de Jong van den Berg L. Consumer reporting: a new step in pharmacovigilance? An overview. Drug Saf. 2003;26:211–7.
44. Gonzalez C, Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Improving ADR reporting. Lancet. 2002;360:1435.
45. Zwillich T. How Vioxx is changing US drug regulation. Lancet. 2005;366:1763–4.
46. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Clin Epidemiol. 2009;62:499–505. doi:10.1016/j.jclinepi.2009.01.012.
47. Taylor R, Bethell H, Brodie D. Clinical trial versus the real world: the example of cardiac rehabilitation. Br J Cardiol. 2007;14:175–8.
48. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. Can Med Assoc J. 2009;180:E47–57. PMC2679824.
49. Tappenden P, Chilcott J, Ward S, Eggington S, Hind D, Hummel S. Methodological issues in the economic analysis of cancer treatments. Eur J Cancer. 2006;42:2867–75.
50. Chilcott J, Brennan A, Booth A, Karnon J, Tappenden P. The role of modelling in prioritising and planning clinical trials. Health Technol Assess. 2003;7(iii):1–125.
51. Ramsdell JW, Braunstein SN, Stephens JM, Bell CF, Botteman MF, Devine ST. Economic model of first-line drug strategies to achieve recommended glycaemic control in newly diagnosed type 2 diabetes mellitus. Pharmacoeconomics. 2003;21:819–37.
52. Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. J Health Serv Res Policy. 1998;3:233–45.
53. Leiter LA, Ceriello A, Davidson JA, Hanefeld M, Monnier L, Owens DR, et al. Postprandial glucose regulation: new data and new implications. Clin Ther. 2005;27 Suppl 2:S42–56.
54. Plosker GL, Figgitt DP. Repaglinide : a pharmacoeconomic review of its use in type 2 diabetes mellitus. Pharmacoeconomics. 2004;22:389–411.
55. What is comparative effectiveness. http://www.effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/. Accessed 6 Aug 2013.
56. Initial national priorities for comparative effectiveness research. The National Academies Press; 2009. http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx
57. The justification for randomized controlled trials. Accessed at http://pluto.mscc.huji.ac.il/~mszucker/DESIGN/pocock.pdf
58. Heeley E, Riley J, Layton D, Wilton LV, Shakir SA. Prescription-event monitoring and reporting of adverse drug reactions. Lancet. 2001;358:1872–3.
59. Lassila R, Rothschild C, De Moerloose P, Richards M, Perez R, Gajek H. Recommendations for postmarketing surveillance studies in haemophilia and other bleeding disorders. Haemophilia. 2005;11:353–9.
60. The future of drug safety. http://www.nap.edu/books/0303103045/html/1.html (2006). Accessed 3 Apr 2007.
61. Drug's black box warning violations in outpatient settings putting patients at risk. http://www.medicalnewstoday.com/medicalnews.php?newsid=37735 (2006). Accessed 3 Apr 2007.
62. DeVane CL, Grasela Jr TH, Antal EJ, Miller RL. Evaluation of population pharmacokinetics in therapeutic trials. IV. Application to postmarketing surveillance. Clin Pharmacol Ther. 1993;53:521–8.

Chapter 6
The Role of the USA Food and Drug Administration in Clinical Research

Stephen P. Glasser, Carol M. Ashton, and Nelda P. Wray

> *Some bargains are Faustian, and some horses are Trojan. Dance carefully with the porcupine, and know in advance the price of intimacy.* [1]

Abstract The USFDA is an agency of the US Department of Health and Human Services and is the nation's oldest consumer protection agency whose function it is to review drugs before marketing, monitor marketed drugs, monitor drug manufacturing and advertising, protect drug quality, and to conduct applied research. It is charged with overseeing of not only human drugs and biologics, but also veterinary drugs, foods, medical devices, and radiopharmaceuticals, and as such serves as a watchdog over industry. This chapter discusses the historical development of the FDA, and what the FDA is today. The phases of research development (phase 0 through phase 5) leading to the marketing of a new drug, the role of the FDA in surgical interventions and medical device approval, and the FDA's role in advertising and adverse event reporting are discussed.

Keywords FDA history • Drug development. Phase 0–3 trials • Futility/feasibility trials. First in man trial • Investigational Drug Application (IND) • New Drug Application (NDA) • International Conference on harmonization (ICH) • Medical device approval • Adverse event reporting • FDA and advertising • Off-label drug use

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

C.M. Ashton, M.D., MPH • N.P. Wray, M.D., MPH
Department of Surgery, Center for Outcomes Research,
Houston Methodist Hospital Research Institute, Houston Methodist
Hospital, 6550 Fannin Street, SM 1661, Houston, TX 77030 USA

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_6,

The United States Food and Drug Agency (USFDA or FDA) is an agency of the US Department of Health and Human Services and is the nation's oldest consumer protection agency, whose functions are to review drugs before marketing, monitor marketed drugs, monitor drug manufacturing and advertising, protect drug quality, and to conduct applied research. The FDA is charged with overseeing of human drugs and biologics; but, also veterinary drugs, foods, medical devices, and radiopharmaceuticals. As such, the FDA serves as a watchdog over industry.

Historical Considerations

The history leading up to the formation of the FDA is interesting. In the early days of our country, epidemics were common (diphtheria, typhoid, yellow fever, small pox etc.), there were few if any specific treatments, the few patent medicines were largely unregulated, some were dangerous, and very few were effective. In fact, drugs readily available in the 1800s would likely astound the average citizen today. For example Winslow's Soothing Syrup and Koop's Babyfriend contained liberal amounts of morphine, a marketed cough syrup contained heroin. Beginning as the Division of Chemistry and then (after July 1901) the Bureau of Chemistry, the modern era of the FDA dates to 1906 with the passage of the Federal Food and Drugs Act; this added regulatory functions to the agency's scientific mission and was the result of recurrent food safety scares (Fig. 6.1). The Division of Chemistry investigation into the adulteration of agricultural commodities was actually initiated as early as 1867. When Harvey Washington Wiley arrived as chief chemist in 1883, the government's handling of the adulteration and misbranding of food and drugs took a decidedly different course, that eventually helped spur public indignation of the problem. Wiley expanded the division's research in this area, exemplified by Foods and Food Adulterants, a ten-part study published from 1887 to 1902. He demonstrated his concern about chemical preservatives as adulterants in the

Fig. 6.1 Depictions of historical developments in drug safety

highly publicized "poison squad" experiments, in which able-bodied volunteers consumed varying amounts of questionable food additives to determine their impact on health. Wiley unified a variety of groups behind a federal law to prohibit the adulteration and misbranding of food and drugs, including state chemists and food and drug inspectors, the General Federation of Women's Clubs, and national associations of physicians and pharmacists [2].

Languishing in Congress for 5 years, the bill that would replace the 1906 Act was ultimately enhanced and passed in the wake of a therapeutic disaster in 1937. In September and October of 1937, in response to consumer demand for a liquid cough medicine, people across the country started dying after drinking a new cough medicine known as Elixer Sulfanilamide (produced by the S.E. Massengill Company). The Elixer was released to the public after testing for flavor, appearance and fragrance—but not toxicity. At the time, federal regulations did not require companies to certify that their drugs were safe, and the solution used to liquefy the sulfanilamide was diethylene glycol, a deadly poison that is found in anti-freeze. From the first death to the FDA's no-holds-barred response, John Swann, in a DVD entitled the ELIXER OF DEATH, tells the remarkable story of the incident that led to passage of the 1938 Food, Drug, and Cosmetic Act, which increased the FDA's authority to regulate drugs. Survivors recall their harrowing ordeals, and FDA historians reveal how agents located 234 of the 240 gal produced—often one bottle at a time!

The public outcry not only reshaped the drug provisions of the new law to prevent such an event from happening again, it propelled the bill itself through Congress. Franklin Roosevelt signed the Food, Drug, and Cosmetic Act on June 25th, 1938. The new law brought cosmetics and medical devices under Federal control, and it required that drugs be labeled with adequate directions for safe use. Moreover, it mandated pre-market approval of all new drugs, such that a manufacturer would have to prove to the FDA that a drug was safe before it could be sold. It irrefutably prohibited false therapeutic claims for drugs, although a separate law granted the Federal Trade Commission jurisdiction over drug advertising. The act also corrected abuses in food packaging and quality, and it mandated legally enforceable food standards. Tolerances for certain poisonous substances were addressed. The law formally authorized factory inspections, and it added injunctions to the enforcement tools at the agency's disposal.

The Bureau of Chemistry's name changed to the Food, Drug, and Insecticide Administration in July 1927, when the non-regulatory research functions of the bureau were transferred elsewhere in the department. In July 1930 the name was shortened to the present version. The FDA remained under the Department of Agriculture until June 1940, when the agency was moved to the new Federal Security Agency. In April 1953 the agency again was transferred, to the Department of Health, Education, and Welfare (HEW). Fifteen years later FDA became part of the Public Health Service within HEW, and in May 1980 the education function was removed from HEW to create the Department of Health and Human Services, FDA's current home. To understand the development of this agency is to understand the

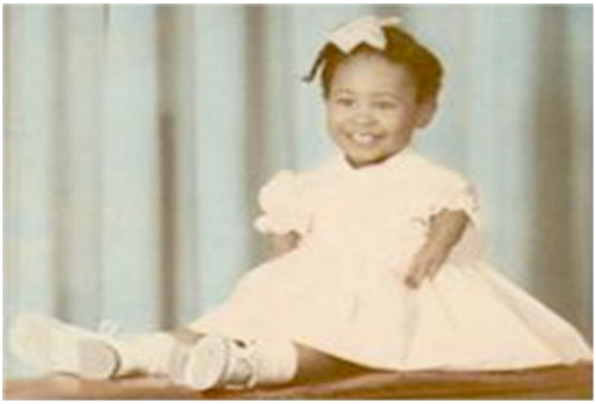


Fig. 6.2 The effects of thalidomide

laws it regulates, how the FDA has administered these laws, how the courts have interpreted the legislation, and how major events have driven all three [3].

During the time period from 1906 and 1938, there were the beginnings of pharmaceutical research and drug discovery. For example, penicillin was discovered in 1928 and insulin was also discovered during this time period. In1951 the Durham-Humphrey Amendment defined the Over the Counter drugs. The 1940–1950s was also the golden age for pharmaceutical companies and over 90 % of all drugs used in 1964 were unknown before 1938. In 1960 the Kefauver Hearings were evaluating drug costs, prices and profits but the 1962 thalidomide tragedy resulted in the 1962 Kefauver-Harris Drug Amendments (Fig. 6.2). The original impetus for the effectiveness requirement was Congress's growing concern about the misleading and unsupported claims made by pharmaceutical companies about their drug products coupled with high drug prices [4]. In this 1962 Act, Congress amended the Federal Food, Drug, and Cosmetics Act to add the requirement that to obtain marketing approval, manufacturers demonstrate the effectiveness (with "substantial" evidence) of their products through the conduct of adequate and well-controlled studies (prior to this amendment there were only safety requirements). This amendment also established informed consent procedures, the reporting process for adverse drug events, and placed drug advertising under FDA jurisdiction. Another important milestone for the FDA came in 1968 when the Drug Efficacy Study Implementation (DESI) was enacted. DESI required that over 4,000 drugs marketed between 1938 and 1962 undergo evaluation for efficacy and safety based upon the existent literature (pre 1938 drugs were "grandfathered"). Other significant actions followed, including the Medical Device Amendment of 1978 that put medical devices under the same kinds of Good Medical Practice (GMP) and Good Clinical Practice (GCP) guidelines that applied to drug development. GCP is an international ethical and scientific standard for designing, conducting, recording, and reporting trials that involve the participation of human subjects. The GCP principles have their origin in the Declarations of Helsinki.

The FDA Now

The U S. Food and Drug Administration is a scientific, regulatory, and public health agency that oversees items that added together account for 20 cents of every dollar spent by consumers on products in the US. Its jurisdiction encompasses most food products (other than meat and poultry), human and animal drugs, therapeutic agents of biological origin, medical devices, radiation-emitting products for consumer, medical, and occupational use, cosmetics, and animal feed. As prior mentioned the agency grew from a single chemist in the U.S. Department of Agriculture in 1862 to a staff of approximately 12,000 employees and a budget of $4 billion in 2012, comprising chemists, pharmacologists, physicians, microbiologists, veterinarians, pharmacists, lawyers, and many others. About one-third of the agency's employees are stationed outside of the Washington, D. C. area, staffing over 150 field offices and laboratories, including five regional offices and 20 district offices. Agency scientists evaluate applications for new human drugs and biologics, complex medical devices, food and color additives, infant formulas, and animal drugs. Also, the FDA monitors the manufacture, import, transport, storage, and sale of over $1 trillion worth of products annually at a cost to taxpayers of over $3 per person. Investigators and inspectors visit more than 16,000 facilities a year, and arrange with state governments to help increase the number of facilities checked.

An era of rapid change for the FDA also occurred with an increase in drug development and approval beginning in the early 1970s. During the period of 1970–2002, reports on the adverse events of over 6,000 marketed drugs numbered in the millions, with 75 drugs removed from the market and another 11 that had severe restrictions placed on their use. From 1975 to 1999, 584 new chemical entities were approved, and over 10 % of these either were withdrawn or received a "black-box" warning. This rapid increase in marketed drugs placed a tremendous burden on the post-marketing safety systems, which the FDA had in place to protect public safety. Subsequently, a number of drug embarrassments' occurred that again reshaped the FDA. These embarrassments included concealed studies (studies that the manufacturer did not publish), fraudulent data (as exemplified in the development of telithromycin-Ketek), rofecoxib (the withdrawal of Vioxx still has pending litigation), and rosiglitazone withdrawal. After rofecoxib was withdrawn from the market, the Center for Drug Evaluation and Research (CDER) asked the Institute of Medicine (IOM) to assess the US drug-safety system. Their report was released in 2006. As to telithromycin, French pharmaceutical company Hoechst Marion Roussel (later Sanofi-Aventis) started phase II/III trials of telithromycin (HMR-3647) in 1998. Telithromycin was approved by the European Commission in July 2001 and subsequently came on sale in October 2001. In the USA, telithromycin gained FDA approval April 1, 2004. FDA staffers publicly complained that safety problems were ignored, and congressional hearings were held to examine those complaints. Some of the data in clinical trials submitted to the FDA turned out to be fabricated, and one doctor went to prison. An indictment said that one doctor fabricated data she sent to the company. Documents, including internal Sanofi-Aventis emails, show that Aventis was worried about this doctor early in study 3014 but didn't tell the FDA until the agency's own inspectors discovered the problem independently [5].

Due to the rapid increase in new drug development during the 1970s and on, The Prescription Drug User Fee Act (PDUFA), was enacted in 1992 and was revised in 1997 and 2002. PDUFA is a program under which the pharmaceutical/biotechnology industry pays certain "user fees" to the Food and Drug Administration (FDA). In exchange for these fees, the FDA agreed, via correspondence with Congress, to a set of performance standards intended to reduce the approval time for New Drug Applications (NDA) and Biological License Applications (BLA). PDUFA assess three types of user fees: fees on applications (NDA/BLA); annual fees on establishments; and renewal fees on products [6, 7]. The law includes a set of "triggers" designed to ensure that appropriations for application review are not supplanted by user fees. These triggers require that Congressional appropriations for such review reach certain levels before user fees may be assessed, and that the FDA devotes a certain amount of appropriated funds annually to drug review activities. However, little provision was made for post marketing drug surveillance. PDUFA resulted in a reduction in review times from 33 to 14 months. Also, prior to PDUFA, the testing ground for new drugs occurred predominantly in Europe. In 1980, only 2 % of new drugs were first being first used in the USA; by 1988 60 % were first used in the USA. The glut of new approved and arguably understudied drugs on the US market, placed a stress on the already inadequate post marketing surveillance systems, and ultimately led to the commission of an Institute of Medicine review. This IOM review led to the FDA Amendments Act of 2007 [8]. This 156 page document expands the authority of the FDA particularly as it relates to marketed drugs (see Chap. 5) Briefly, this new act grants the FDA the power to require postmarketing studies, to order changes in a drug's label, and to restrict distribution of a drug. The Act also provides new resources (225 million dollars over 5 years) aimed at improving drug safety.

International Conference on Harmonization (ICH)

Ultimately, an international effort was initiated that was designed to bring together the regulatory authorities of Europe, Japan and the United States and experts from the pharmaceutical industry in the three regions to discuss scientific and technical aspects of product registration. Their stated purpose is to make recommendations on ways to achieve greater harmonization in the interpretation and application of technical guidelines and requirements for product registration in order to reduce or obviate the need to duplicate the testing carried out during the research and development of new medicines. The objective of such harmonization is a more economical use of human, animal and material resources, and the elimination of unnecessary delay in the global development and availability of new medicines while maintaining safeguards on quality, safety and efficacy, and regulatory obligations to protect public health. The Mission Statement of the ICH (as taken from their website) is "*to maintain a forum for a constructive dialogue between regulatory authorities and the pharmaceutical industry on the real and perceived differences in the technical requirements for product registration in the EU, USA and Japan in order to ensure a more timely introduction of new medicinal products, and their availability to patients; to contribute to the protection of public health from an international perspective; to monitor and update harmonized technical requirements leading to a greater mutual acceptance of research and development data; to avoid divergent future requirements through harmonization of selected topics needed as a result of therapeutic advances and the development of new technologies for the production of medicinal products; to facilitate the adoption of new or improved technical research and development approaches which update or replace current practices, where these permit a more economical use of human, animal and material resources, without compromising safety; and, to facilitate the dissemination and communication of information on harmonized guidelines and their use such as to encourage the implementation and integration of common standards*." [9].

USA Drug Development Phases

Since the FDAs1962 amendment mentioned above, the FDA, Industry, and academia have debated the issue of what constitutes "*sufficient evidence of effectiveness*". Before getting to that point, there is a fairly regimented program for drug development that will be discussed in the following paragraphs.

Preclinical Evaluation

First, when a new molecule is identified as a possibly active drug, it undergoes chemical and physical characterization and screening for biological activity by testing in appropriate animal models. This includes toxicity studies followed by preclinical pharmacology where dosage, mode of action, chronic toxicology, safety, efficacy, and teratogenicity are evaluated. If the drug seemingly has merit, it advances to clinical investigation where it undergoes three phases of evaluation (phase 1, 2 and 3 trials). Recently, a new phase, phase 0, has been added, especially for cancer drug development. However, before clinical testing can take place, an Investigational New Drug (IND) Application must be submitted and approved by the FDA. Since across state transfer of drugs is necessary for most drug related clinical research, and there is a federal law against such transport, an IND allows for an exemption in the law so that a drug can be shipped via interstate commerce. This is a rapid process (the FDA must respond to the IND application within 30 days). Parenthetically, the FDA uses a broad definition for "new drug". It is not just a new chemical moiety; rather a new drug is any drug or drug use that is not

Table 6.1 Uses of phase 0 trials

Determine if a MOA defined in non-clinical models applies to humans
Provide PK/PD data before phase 1 trials
Evaluate PK/PD of analogs in order to select the most promising candidate
Determine a dose range
Refine biomarker assays
Develop novel imaging probes and evaluate its physiology in humans

included in current labeling of that drug. If the drug has no prior approval the definition is fairly obvious. However, an approved drug now being studied with a new release system (e.g. a transdermal patch, or a new salt side chain), a new indication, or a new combination (even if the two separate drugs are approved) is considered “new”. So, for example, when aspirin was to be studied in the Coronary Drug Project, an IND had to be submitted for this “new” drug [10].

Phase 0–3 Studies

Following IND approval, the new phase 0 clinical trial was developed in response to the FDA’s recent exploratory Investigational New Drug (IND) guidance predominantly for the study of new cancer drugs (Table 6.1). Even though phase 0 studies are done in humans, phase 0 studies are exploratory studies that often use only a few small doses of a new drug in each patient, to test whether the drug reaches the tumor, how the drug acts in the human body, and how cancer cells in the human body respond to the drug. The patients in these studies might need extra biopsies, scans, and blood samples as part of the study. The biggest difference between phase 0 and the later phases of clinical trials is that there’s no chance the volunteer will be helped by taking part in a phase 0 trial. Because drug doses are low, there’s also less risk to the patient in phase 0 studies compared to phase I studies. Phase 0 studies are not yet being used widely, and there are some drugs for which they wouldn’t be helpful. Phase 0 studies are very small, often with fewer than 20 people. They are not a required part of testing a new drug, but are part of an effort to speed up and streamline the process [11].

More commonly the investigation of a new drug begins with a phase 1 study and, these are more commonly referred to as “first-in-man” studies. In general phase 1 trials have relatively small sample sizes and are usually performed in normal human volunteers. The goal is to evaluate pharmacokinetics (PK) and to determine if there are any differences compared to the preclinical studies. Early, phase 1 studies are acute PK evaluations; later the studies may include chronic PK and dose escalation in order to determine the maximum tolerated dose. First in man studies have received renewed interest partly as a result of the TGN-1412 study, which in its first human clinical trials, caused catastrophic systemic organ failure in the subjects, despite

Table 6.2 Null hypotheses for different study questions

	Null hypothesis	Alternative hypothesis
Superiority	$\text{New} = \text{Old}$	$\text{New} \neq \text{Old}$
Equivalence	$\text{New} < \text{Old} + \delta$	$\text{New} = \text{Old} + \delta$
Futility	$\text{New} > \text{Old}$	$\text{New} \ngtr \text{Old}$

being administered at a supposed sub-clinical dose [12]. The adverse events in this trial resulted in the hospitalization of six volunteers. At least four of the six these suffered multiple organ dysfunction, and one trial volunteer was said to be showing signs of developing cancer. Tentative opinions from an as yet uncompleted inquiry suggest that the problems arose due to an "unforeseen biological action in humans", rather than any breach of trial protocols; and, the case, therefore, has had important ramifications for future trials of potentially powerful clinical agents. In part, as a result of this aforementioned trial, the European Medicines Agency (EMEA the European equivalent of the USFDA) in 2007 approved draft guidelines for first in man studies [13]. This initial draft guidance has been the subject of wide comment in the clinical trials community, and as a result of a wide variety of opinions has been a challenge to finalize the guidelines [14]. Additionally, there has been some discussion about whether it is more appropriate to use healthy volunteers or patients in these first-in-human trials [15]. Healthy volunteers are preferred if the perceived risk is low, but patients afflicted with the target disease might be appropriate when there is potential therapeutic benefit. However, the advantages of normal human volunteers for the questions to be answered by these phase 1 trials include greater subject homogeneity, less confounding by co-morbid conditions and other drugs the subject might be taking, and less confounding of drug effects by symptoms, signs, and laboratory findings.

Phase 2 trials are slightly larger and also examine PKs, but now in patients with the disease of interest. In addition, these are referred to as "proof of concept" studies. They are also dose-ranging, feasibility, futility, and safety studies. Recently, there has been a suggestion that phase 2 trials be sub-classified into 2A and 2B. Phase 2B studies can be thought of as smaller early RCTs, while phase 2A studies are an extension of phase 1 studies, but in patients rather than subjects. These classifications are not firm, however, and there are many exceptions. Phase 2 studies can also be feasibility studies, in which efficacy, response rates, and response durations are determined. This is also a phase in which ineffective treatment can be rejected (futility study) prior to the more expensive phase 3 trials (since there are an increasing number of phase 3 efficacy trials which fail to find benefit). In this regard, in order to save money and time, phase 2 futility studies are becoming more common. In this variant of phase 2 trials, futility studies can be designed as a way of dealing with the trade-off between investment risk and clinical promise. That is, one way to reduce the studies sample size is to focus on futility-that is designing a study to identify which agents are least likely to demonstrate benefits rather than the more typical goal of identifying the most promising agents. The null hypothesis in a futility study is that the treatment has promise and will therefore produce results exceeding a meaningful threshold (Table 6.2). If that threshold is not met, the null is rejected and further study is considered futile. Remember, the same provisos hold regarding the null discussed in chapters 3 and 18. That is, agents passing an efficacy criterion are winners, but agents meeting the futility criterion are merely non-losers. Palesch et al. evaluated 6 phase 2 futility study designs of therapeutic stroke trials. They identified 3 trials as futile in phase 2, and none of the 3 subsequently showed benefit in phase 3 trials. In the remaining 3 phase 2 trials that did not show futility, 1 showed efficacy in phase 3 [16].

More specifically, the way phase 2 futility studies are designed is first to estimate the proportion of favorable outcomes in untreated controls (this is usually done from historical, case-series, or control groups from previous trials) and this becomes the proportion of favorable outcomes for the single arm phase 2 futility study. The minimally worthwhile improvement of the drug under study is then estimated as one does in determining the sample size in phase 3 studies. If the null hypothesis is rejected that there is a minimally worthwhile improvement, we conclude that the benefit of treatment is less than what we would want, and it is therefore futile to proceed to a phase 3 trial. Additionally, in phase 2 futility trials, one would want to minimize the risk of drawing false negative conclusions (that is the study suggests that the drug has no efficacy when it in fact does-one would not want to miss studying a potentially effective agent). The sample size is then "hedged" towards this aforementioned goal, with less concern about a false positive conclusion (that is that the drug is effective when in fact it is not) [16].

Phase 3 trials are classical efficacy studies generally using RCT designs as discussed in Chap. 3; and, phase 4 studies are discussed in Chap. 5. However, it is the phase 3 study that was the topic of discussion resulting in the 1962 Kefauver-Harris amendment. During those hearings, the main issue of contention about phase 3 studies surrounded the words "*substantial evidence*" of effectiveness that the FDA required for drug approval. In the above mentioned FDA Act, substantial evidence was defined as "*evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could be fairly and responsibly concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof.*" The argument that ensued from this definition centered on what the specific quality of evidence was in order to establish efficacy. It was the FDA's position that Congress intended to require at least 2 adequate and well-controlled studies, each convincing on its own, to establish efficacy. There has been some subsequent flexibility by the FDA in regard to the above as it applies to a specific drug in development. In some cases, for example, the FDA has relied on information from adequate and well-controlled studies published in the literature. In other cases where it would be difficult to perform a second study due to ethical concerns, the result of a single study could be accepted (as long as it was of excellent design, provided highly reliable and statistically strong – $p < .001$ – with evidence of important clinical benefit-such as survival). The phases of drug development are summarized in Table 6.3.

Table 6.3 Contrasts between the developmental pathways of new prescription drugs and new invasive therapeutic procedures

Phase or stage	Drug approval process	Invasive procedure process
0	Preclinical experiments (animal or bench)	Proof of principle: usually a report of a case or small case series in patients that describes the technique
1	Single-group trial (n = 20–80 volunteers)	Refinement and definition: modification of the technique from early experience… this phase is often unreported
2	Controlled trials in several hundred people with the target condition	Dissemination: technique is adopted rapidly by other surgeons who then report their case series
3	Controlled and single group trials for the purpose of establishing benefit:risk ratio	Comparison with current standard approaches: The technique has stability and popularity but is it better than other treatment? Should be answered by a RCT but often isn't
4	Post FDA approval: Post marketing studies	Surveillance and quality control: monitoring of complication rates

From McCulloch [17]

The requirement of more than 1 adequate and well-controlled investigation reflects the need for independent substantiation of experimental results and refers back to the question posed in Chapter 3 that asked why studies could presumably be of similar design and yet lead to different results. Indeed, the FDA realized that any clinical trial might be subject to unanticipated, undetected, systematic biases that may be operative irrespective of the best intentions of sponsors and investigators. They also note that the inherent variability in biological systems may produce a positive trial by chance alone. In addition, results may be dependent on specific issues related to the site or the investigator (e.g. concomitant treatments, diets etc.) that may impact the generalizability of the results. Finally (and fortunately rarely), favorable efficacy might be the product of scientific fraud. Independent substantiation of experimental results then addresses these problems by providing consistency across more than 1 study, thus greatly reducing the possibility that a biased, chance, site-specific, or fraudulent result will lead to an erroneous conclusion that a drug is effective.

The concept of independent substantiation of trial results has often been referred to as replication, but replication may imply precise repetition of the same experiment. Actually, studies that are of different design, in different populations, with different endpoints or dosage forms may provide evidence of efficacy, and this may be even more convincing than repetition of the same study. It should be noted, that it is usually not necessary to rely on a single study to support the efficacy of a drug under development. This is because, in most situations there is a need to explore the appropriate dose range, to study patients with differing complexities and severities of disease, to compare the drug to other therapies, to perform safety studies, so that before marketing, most drugs will have been evaluated in more than 1 study.

Another trend seen by the FDA is the increase in new drug applications that use foreign studies as the basis for approval. In 2000, 27 % of NDA's contained pivotal data from foreign studies [18]. There is no current restriction on non-US studies being used to support an NDA so long as they are well designed and conducted and the study sites are available for inspection. This trend regarding the observation that "the frequency with which initial clinical trials were being performed outside of the US" has been discussed by DeMaria [19]. His concerns centered around the disadvantages of conducting a clinical investigation outside the sponsors country (e.g. logistic and language differences) and include: monitoring of such sites; the apparatus needed for clinical research may be less developed and investigators less experienced; the nature and disease etiologies of the patients may differ from the US; and genetic or genomic differences may exist. He also raised concern as to "*whether we in America are exploiting the rest of the world to prematurely test potentially hazardous therapies, or conversely, whether our regulatory and financial environment is stifling access to important new innovations for patients and investigators.*" DeMaria does note the many advantages of performing clinical research in other countries noting the potential importance of the globalization of clinical research. Advantages he mentioned include: the cost of conducting studies and the ability to recruit patients into studies might be greater in other countries; and, that in other countries regionalization of health care facilities, that concentrates patients, is greater in many countries vs. the US. These trends are also a result of a decline in the number of active investigators in the US (a decline of 3.5 % annually since 2001 compared to a 13.5 % increase outside the US).

Over recent years, there has been increasing concern about industry-funded bias. Currently the pharmaceutical industry funds overall more than half of the research done today, while the NIH accounts for 29 %. Some have observed that industry funded studies may be 4–5 times more likely to be favorable to the sponsors drug than non-industry sponsored studies [20–22]. In a Point/Counterpoint editorial Jeff Steri argues 'that it is not the source of research funding that counts: it is the quality of the research that matters" [23]. He also points out that we all have biases regardless of the source of funding. At least in the nutrition research arena the source of funding did not affect the quality of the research [24]. (See further discussion in Chap. 19).

Another potential problem relates to the lack of reporting the results of clinical trials in general and drug development clinical trials more specifically, particularly if they are negative. For drugs that receive FDA approval, disclosure of trial results can occur in a number of ways, including the FDAs own Summary Basis of Approval; but there are limitations to that reporting. Lee et al. conducted a study of trials supporting new drugs approved between 1998 and 2000 and determined their publication status and time from approval to full publication in the medical literature at 2 and 5 years [25]. Of 909 trials only 43 % were published after 5 years, although for trials classified as "pivotal", 76 % were published, and in both cases publication favored trials that were positive. Perhaps surprising to some, this lack of publication also applied to NIH funded trials [26]. In review, a number of clinical trial facts pertain to the FDA approval process as listed in Table 6.4.

Table 6.4 Clinical trial facts

Clinical trial facts
In 2004, between five and six million people participated in some 80,000 clinical research studies in the U.S., and more will likely be needed in coming years to fulfill safety and efficacy requirements
The top 20 drug companies spend $30 billion on research and development, about 40 % of which goes to fund clinical trials
Eight of the top 15 drug companies did not get the go-ahead for a single drug last year
In 1980, drug companies spent some $2 billion on R&D, and 34 new drugs were approved. In 2000, they spent close to $30 billion, but only 24 drugs were approved
Seven of ten of drugs approved by the FDA never make enough money to justify their development costs
Completing all the phases of clinical trials required for approval of a new drug can cost anywhere from $300 to $800 million
Although the NIH budget has doubled in the past 5 years—with the implied purpose of encouraging the development of new drugs—the FDA's budget remains inadequate to review these drugs for qualification
Drug costs are increasing by about 18 % a year, but only 4 % is due to price increases. The rest is the result of replacing older, more invasive, expensive, and less effective medical treatments

FDA and Medical Devices

When the US Congress first mandated in 1938 that medical products demonstrate safety and effectiveness, the law applied only to drugs. It was not until 1976 that the law was amended to include devices (mostly the result of problems related to intrauterine devices). The 1976 law included the need for premarket approval of medical devices that was similar to that for new drugs. However, thousands of devices were already marketed in 1976 so an alternative pathway [termed the 510(k) provision] from the more rigorous premarket approval process was added to the law, in order to enable newer versions of existing devices to enter the market. This alternative process did not require clinical trials but only required that the manufacturer demonstrate that the device was substantially equivalent in materials, purpose, and mechanism to a device already on the market (Fig. 6.3). The potential problem with this alternative pathway was demonstrated by Zuckerman et al. who found that of the 113 recalls from 2005 to 2009 that the FDA determined could cause serious health problems or death, 71 % had been cleared by the alternative 510(k) process [27]. Devices are classified as to their perceived risk using a 3-tierd system: Class I devices are thought to be low risk; Class II, higher risk, but where substantially equivalent to existing approved devices might just require bench and/or animal testing; and Class III devices are either life sustaining/supporting, or present a high risk of illness or injury (for example heart valves, pacemakers, etc.) (Table 6.5). The standards for demonstrating safety and effectiveness are determined in part by this classification [28]. Women particularly are underrepresented in terms of safety

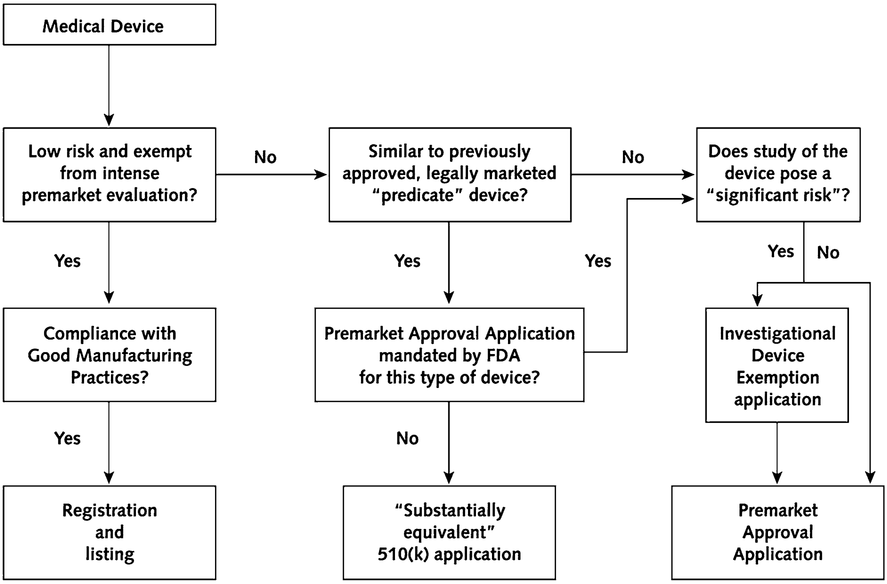

Medical Device
Low risk and exempt from intense premarket evaluation?
No
Similar to previously approved, legally marketed "predicate" device?
No
Does study of the device pose a "significant risk"?
Yes
Yes
Yes
Yes
No
Compliance with Good Manufacturing Practices?
Premarket Approval Application mandated by FDA for this type of device?
Investigational Device Exemption application
Yes
No
Registration and listing
"Substantially equivalent" 510(k) application
Premarket Approval Application

Fig. 6.3 Overview of the medical device approval process (From Maisel [30])

Table 6.5 Classes of medical devices relative to the degree of testing

Class I	Devices that are thought to be low risk
Class II	Higher risk, but substantially equivalent to existing approved devices and might just require bench and/or animal testing
Class III	Devices are either life sustaining/supporting, or present a high risk of illness or injury (for example heart valves, pacemakers, etc.)

and effectiveness data in trials of CV devices according to Dhruva et al. [29]. The Institute of Medicine's report examining the impact of changes in government support for women's health research recommended that "all medical product evaluations by the Food and Drug Administration present efficacy and safety data separately for men and women." [30] Dhruva et al. performed a systematic review of the demographics, comments on gender bias, and analysis of results by sex for 78 high-risk cardiovascular devices that received premarket approval by the FDA between 2000 and 2007. FDA summaries of evidence did not report sex of enrollees in 34 (28 %) of 123 studies. For studies reporting sex distribution, the study populations were, on average, 67 % men. There was no increase in the enrollment of women over time.

As with the drug approval process, the postmarket evaluation of medical devices was also thought to be wanting, and in 1986, 400 patients were affected by a mechanical valve strut failure, resulting in the Safe Medical Devices Act of 1990 and the Medical Device Amendments of 1992 which strengthened the FDAs authority in monitoring post market surveillance [31].

There are similarities in the regulatory process in the US compared to the European Union (EU), but differences exist. For example the EU relies heavily on independent commercial organizations to implement regulatory oversight of medical devices (these organizations are called notified bodies or NBs). Also, to receive approval to market class III and some class II devices in the EU, the manufacturer must demonstrate that the device is safe and performs in the manner consistent with the manufacturers intended use, while in the US prospective RCTs are usually necessary [32].

FDA and Surgical Interventions

Carol M. Ashton and Nelda P. Wray
Co-Director, Department of Surgery, Center for Outcomes Research,
Houston Methodist Hospital Research Institute, Houston Methodist
Hospital, 6550 Fannin Street, SM 1661, Houston, TX 77030, USA

As already discussed, prescription drugs are regulated by the FDA, and for marketing approval of a new drug there is the requirement that there be pre-release demonstration in randomized trials of its efficacy and safety in humans. Medical devices, including the assistive and implantable devices used in surgical procedures, are also regulated by the FDA, though the pre-market evaluative process is less stringent than it is for prescription drugs; most devices are cleared for marketing by the agency without the requirement of human tests of device safety and effectiveness. The FDA's regulatory purview does not extend to surgical procedures, and there are no FDA regulations governing surgical interventions. Rather, new surgical interventions are developed based on anatomic and clinicopathological correlations in humans and studies in animals, and then used in humans, with the initial experience reported as case reports or a series of cases. Subsequent large-scale dissemination of the procedure occurs as additional surgical groups begin using it. In most cases, it is only when doubts set in about a given procedure that its efficacy is evaluated in a randomized controlled trial. The contrasts between the developmental pathways of new prescription drugs and new surgical procedures are shown in Table 6.6.

The reasons behind the existence of a double standard for pre-release proof of effectiveness and safety—one for prescription drugs and biologics and another, much less evidence-based, for surgical procedures—are numerous and beyond the scope of this chapter [33]. But settling for less rigorous evidence, namely that generated by case series, has consequences, as can be seen from instances in which a surgical procedure has been tested using the gold standard study design for efficacy tests, the randomized controlled trial. Those RCTs generally demonstrate that the procedure is less beneficial or more harmful than originally thought, no better than a nonoperative course of action, beneficial for only certain subgroups, or no better than a placebo (sham procedure).

Table 6.6 Different designs for surgical studies: strengths and weaknesses

Comparator	Non-operative therapy	Alternative invasive procedure	Sham procedure
Random allocation[a]	Yes	Yes	Yes
Blind patients[b]	No	Yes	Yes
Blind adjudicator outcomes	No	Sometimes	Yes
Minimizes crossovers	No	Yes	Yes

[a]Controls selection bias
[b]Controls expectancy bias

A classic example is the story of lung volume reduction surgery (LVRS) for emphysema [34]. The first report of the use of LVRS in humans was published in 1957 but the procedure did not become widely used until it was modified in the mid 1990s by Joel Cooper [35, 36]. Dr. Cooper reported his experience with 20 cases in 1994 (abstract) and 1995 (paper). By 1996, 1200 LVRS were performed in Medicare beneficiaries, at an estimated cost of $30,000 to $70,000 each, not counting physician charges. But here is where the LVRS story diverges from the typical scenario. Scrutiny of LVRS by a consensus of experts as well as Medicare officials led to concerns about the procedure's effectiveness and safety. In a landmark decision, Medicare officials decided that coverage for LVRS would only be provided in the context of a clinical trial [37]. Dr. Cooper and others complaining about this decision as unethical because of the "*obvious benefit of the procedure*" challenged this decision. In record time, the NIH, Health Care Financing Administration (now the Centers for Medicare and Medicaid Services) and the Agency for Healthcare Research and Quality launched a randomized trial of LVRS vs. medical therapy for severe emphysema, the National Emphysema Treatment Trial, enrolling the first patient in 1997. The initial results, reported in 2003, indicated that in 1,219 patients followed for an average of 29 months, in certain subgroups of patients, LVRS resulted in higher mortality rates than medical therapy [38]. Based on the trial results, Medicare officials limited coverage to patient subgroups that appeared to benefit or at least not be harmed by LVRS. But the trial seems to have quenched demand for LVRS. By 2006, as reported in the New York Times, "Medicare says it will pay, but patients say 'no thanks,'" only 458 Medicare claims for LVRS were filed between January 2004 and September 2005 [39]. Without the findings of the National Emphysema Treatment Trial, many more patients would have undergone a major operation from which they had no chance of benefit and only a chance for harm.

Two other examples of the consequences of this "evolutionary pattern" in the development of surgical interventions are provided by carotid artery endarterectomy for stroke prevention and arthroscopic treatment for relief of knee pain due to osteoarthritis. The first case report of carotid artery endarterectomy in a human appeared in 1956 [40]. By 1971, 15,000 carotid endarterectomies were performed in the USA. By 1985, this had increased to 107,000 [41]. Criteria were then developed for the appropriate use of this procedure; when they were retrospectively applied to the carotid endarterectomies performed on Medicare beneficiaries in 1981, only 35 % of patients were found to have undergone the procedure for "appropriate" reasons, and in another 32 % the reasons were equivocal [41]. Definitive randomized trials of carotid endarterectomy were not conducted and reported until the mid-1990s [42–44]. The results of the trials changed clinical practice: based upon the appropriateness criteria, by 1999 only 8.6 % of carotid endarterectomies could be deemed "inappropriate" [45]. On the other hand, 75 % of all carotid artery endarterectomies are now performed in asymptomatic patients, in whom the risk:benefit ratio of the procedure is much narrower. In 2004, the FDA approved for use the first carotid artery stent and stent deployment system. Unfortunately, over the past decade numerous observational and interventional studies have provided equivocal results in most patient subgroups about the incremental benefit of stenting over endarterectomy. In the Carotid Revascularization Endarterectomy vs. Stenting Trial (CREST) it was found that among patients with symptomatic or asymptomatic carotid stenosis, the risk of the composite primary outcome of stroke, myocardial infarction, or death did not differ significantly in the group undergoing carotid-artery stenting and the group undergoing carotid endarterectomy. During the periprocedural period, there was a higher risk of stroke with stenting and a higher risk of myocardial infarction with endarterectomy [6, 46].

A final example of the consequences patients and society are paying because of the typical pattern of the dissemination of surgical innovations is that of arthroscopic lavage with or without debridement for knee pain due to osteoarthritis. Fiberoptic arthroscopic debridement for this condition began to be used in the mid-1970s. By 1996, more than 650,000 of these procedures were performed in the US [47]. A definitive randomized trial of the efficacy of this procedure was not begun until 1995. That trial was a single site study in which 180 people were randomized in the operating room to arthroscopic lavage, arthroscopic lavage plus debridement, or a sham procedure (skin incisions with no entry into the joint) and followed for 2 years. The study showed that arthroscopic lavage with our without debridement was no better than the sham procedure in relieving pain and restoring function [48]. Two subsequent trials have confirmed that arthroscopic surgery for knee pain due to osteroarthritis is no better than nonoperative therapy [49, 50]. It is certainly feasible, if challenging, to design and conduct rigorous randomized trials testing the benefits and harms of surgical interventions, and many high-quality surgical trials have been done. However, a large proportion of surgical trials have serious defects in design and/or reporting that undermine their internal and external validity as well as their clinical usefulness [51]. Coupled with the fact that many surgical innovations disseminate without being tested in randomized trials, the poor methodological quality of many of the surgical trials that *are* conducted means that the evidence base for many of the invasive therapeutic procedures in use today is seriously deficient and lags far beyond what we know about prescription drugs.

What are some of the challenges in designing an RCT to evaluate the efficacy of an invasive therapeutic procedure? Potential randomized designs that could be used to evaluate the efficacy of a procedure include comparing the operative procedure to a non-operative course of therapy, the operative procedure against a sham or placebo procedure, and the operative procedure against an alternate operative procedure. Evaluating an operative intervention against a non-operative comparator is by far the most commonly used design, but blinding as to group assignment is impossible, and expectancy bias on the part of patients and outcome assessors can affect estimates of treatment effect, especially if the surgical procedure is intended to alter subjective endpoints such as symptoms or function rather than more objective endpoints, e.g., death rates. In addition, because of participants' and doctors' treatment preferences, crossovers may be a serious problem. For example, in a 2006 RCT of diskectomy vs. nonoperative therapy for lumbar disk herniation, only 60 % of people randomized to surgery actually had the surgery, while 45 % of those randomized to the nonoperative arm crossed over and had the surgery [52]. The use of a sham procedure as a comparator in an RCT is limited, among other things, by the risks associated with sham anesthesia and a sham procedure. These are dictated by the nature of the active invasive procedure that is under evaluation. For many procedures, it would be impossible to design a sham that would maintain blinding yet still be safe for the patient, which is why sham comparators are infrequently used. However, there have been studies that have used a Sham procedure for Parkinson's Disease and not surprisingly this resulted in considerable debate about is appropriateness. In fact, a survey of leading Parkinson disease researchers in the US and Canada found that 97 % believe sham surgery is necessary for evaluating the safety of cell- and gene-based neurosurgical interventions, and fully half of the investigators said that it would be unethical *not* to use sham procedures to test promising cell-based and gene therapies for Parkinson's Disease due to the risk of false positive findings. Even medical ethicists are divided on the issue of sham surgery—some have defended it while others have been outspokenly against the procedure [53]. Finally, comparing a surgical innovation to an invasive procedure that is part of the accepted standard of care is informative only in instances when we are certain about the efficacy of the comparator procedure. Blinding as to treatment group assignment is possible with the latter design, as it is with sham procedure controls. As Baruch Brody has said regarding the issue of blinding in invasive intervention trials, one needs a "*…balancing of the scientific gains from blinding against the burdens imposed on the subjects and deciding when the burdens are too great*" [53]. Table 6.6 summarizes the limitations of each of the above approaches.

Invasive therapeutic procedures pose other challenges in the design of randomized trials to evaluate their efficacy, including but not limited to:

- The need to refine the surgical technique in humans: implications for the timing of RCTs
- Learning curves of individual surgeons
- Unequal technical skill in the individual surgeon for various procedures
- Patient—and doctor—preferences for operative vs. nonoperative intervention
- Clinical uncertainty and equipoise: who defines these? When is a trial justifiable?
- Modest effect sizes expected from most therapeutic interventions and implications for sample size and number of participating surgical centers
- Difficulty of evaluating effects of an intervention aimed at alleviating subjective parameters such as pain, discomfort, disability, etc.
- Placebo effect associated with invasive therapeutic procedures, and
- Control of expectancy bias in outcome assessments (blinding of patient, surgeon, outcome assessors)

In summary, the current standard of practice is that invasive therapeutic procedures are devised and become widely used in the public without first having been put to scientifically valid demonstrations in humans (i.e., randomized controlled trials) to determine the extent to which their benefits exceed their harms and costs and those of alternative courses of therapy. Compared with pre-release standards for prescription drugs, those for invasive procedures seem antiquated at best and do not seem to be serving the interests of patients and our society as well as they could be. As Wennberg stated, "*we need a way to assure the American people that the needed evaluations of clinical theory are done in a timely way, before plausible but wrong ideas get institutionalized into the everyday practice of medicine*" [54]. We need to develop efficient approaches to assessing the benefits and risks of surgical innovations, approaches that yield high-quality, rigorous data while at the same time protecting patients' ethical rights and our society's justifiable interests in innovation and technological advances.

Adverse Event Reporting

Up to now we have discussed the industries role in drug development, and its lack of a role in surgical procedure development. From the FDA standpoint, the interest in monitoring the trials as they proceed and to ensure patient safety during the process is tantamount. Thus, for each trial, a mechanism must be in place for a timely review of adverse events. In fact, one FDA report cited the failure to report adverse events as required as one of the top ten problems surrounding clinical trials. The FDA definition of an adverse event is "*any unfavorable and unintended sign, symptom, or disease temporally associated with the use of a medical treatment or procedure regardless of whether it is considered related to the treatment or procedure*."

The FDA has classified adverse drug events (ADEs) as serious when death, life threatening occurrences, hospitalization, persistent or permanent disability, or the need for medical or surgical intervention occurs during (and up to 30 days after) a clinical trial. An example of this is the report by Suntharalingam et al. that occurred during a phase 1 trial. They describe the events that occurred when six healthy volunteers received a dose of TGN1412 (a monoclonal antibody that affects T-cells). In all six subjects, a life threatening cytokine-release syndrome developed [12].

There are a number of questions that address adverse event reporting as follows:

Are clinical trials powered in such a way as to address differences in ADE's vs. placebo or active control?

The answer to this is generally no. Phase 1–3 trials are powered based on presumed efficacy beyond that of the control treatment, not based upon any ADE frequency. Also, the entire drug development portfolio submitted to the FDA for drug approval may consist of fewer than 5,000 patients exposed and certainly fewer than 10,000. Most of those patients are represented by phase 3 trials, and by the time a phase 3 trials is launched common ADE's will have already been ascertained. Given this, ADE's that occur even at a rate of 1 in 10,000 will not be revealed in

Table 6.7 Strengths and limitations of different ways of ascertaining ADEs

Study design	Strengths	Limitations
Voluntary reporting	Detects signals of rare diseases	Risk difficult to quantify, details incomplete, selective reporting,
RCTs	Reduced confounding	Power, duration may be inadequate
Non-randomized trials	Effectiveness, larger sample, can explore interactions with disease and drugs	Confounding, data based upon computerized records rather than real use
Meta-analyses	Greater sample	Relies on quality of primary data, missing and unreported data

the drug development portfolio submitted to the FDA for drug approval. Table 6.7 summarizes the strengths and limitations of some of the ways of ascertaining adverse drug events.

Does the manner in which ADE's are ascertained matter?

This is a frequently argued point in which there is insufficient information to come to a meaningful conclusion. Of course, most studies report ADE frequency, but the absolute frequency depends upon whether ADE's are ascertained verbally either by general questions (e.g. "have you had any new symptoms since the last visit" or specifically, e.g. "have you had any headaches since the last visit?"); or ascertained by checklists either filled out by the patient or elicited by the study coordinator and/or the principal investigator. One of the attempts to evaluate the differences in the manner of ascertainment comes from the Acute Myocardial Infarction Study (AMIS) as shown in Table 6.8 [55]. Not surprisingly, compared to controls, the frequency of GI bleeding elicited by specific questions was greater than those that were volunteered observations, but the relative difference between the active and control treatments was nearly the same. However, the few studies that have specifically addressed the accuracy of self-reported medical events suggest relatively poor agreement between those reports and medical record review [56, 57]. More recently, Bolland et al. conducted a systematic analysis of the relationship between self-reported and adjudicated events (including unreported events) from a 5 year calcium supplementation study [58]. They found that almost half (48 % of MIs and 42 % of strokes) the events could not be verified; and, 43 % of verified MIs and 10 % of verified strokes were unreported. The authors concluded that although other reports have demonstrated a higher agreement between self-reported events and adjudicated events, if accuracy of a clinical event in a trial is critical, self-reported events should not be relied on.

Does the use of surrogate endpoints affect the determination of ADE frequency?

Recall that a surrogate endpoint is an outcome used in lieu of the real outcome of interest, and the main reason surrogate endpoints are used is so the clinical trial will be of shorter duration and/or can be conducted with a smaller sample size. It is thus obvious that this would decrease one's ability to uncover infrequent ADE's. Also, whereas the FDA would prefer to approve drugs on the basis of a clinically

Table 6.8 Comparison of type of reporting ADEs

% Reporting Selected ADE's in AMIS			
Volunteered	Hematemesis	Tarry stools	Bloody stools
Aspirin	.27	1.34	1.29
Placebo	.09	.67	.45
Elicited			
Aspirin	.62	2.81	4.86
Placebo	.27	1.74	2.99

Whereas the absolute % differs for volunteered vs. elicited, the delta is similar
From: The AMIS Study Group [55]

relevant endpoint, it does consider drugs tested against surrogate endpoints. Between 1998 and 2008, the FDA approved 69 applications for new molecular entities based upon surrogate endpoints [59]. Surrogate endpoints are more fully discussed in Chap. 3.

Does the use of intention-to-treat analysis affect the determination of ADE frequency?

As with the use of surrogate endpoints, ITT analysis can reduce one's ability to determine the true ADE frequency. This is because, if a patient drops out from a trial before completion, and does not receive the drug for the entire trial duration, they will not have been exposed to the drug under study for the full trial time period. Even if they are dropped early in the study for an ADE they might have had an additional ADE (or a more severe ADE that the one that caused them to be withdrawn from the study early), had they been able to continue for the entire study. Since ITT is the primary analysis of a RCT (already a relatively short trial for the reasons mentioned in Chap. 3) most RCTs underestimate the true ADE frequency.

The FDA and Advertising

The FDA has a clear mission of protecting the public health by assuring the safety, efficacy, and security of human drugs. The FDA is also responsible for advancing the public health by helping to speed innovations that make medicines more effective, safer, and more affordable [60]. If we consider that the FDA is also responsible to help the public get accurate, science-based information that is needed for medicines to improve their health, then it is understandable that a key role of the FDA is as a regulator and supervisor of manufacturer promotional activities.

The Division of Drug Marketing and Communications (DDMAC) in the Center for Drug Evaluation and Research, at the US Food and Drug Administration (FDA), is responsible for reviewing sponsor promotional materials, including prescription drug advertising, promotional labeling, and materials prepared for prescribers [61]. The main objective of the division is to ensure that information about prescription drugs disseminated by sponsors to health care providers and consumers is not false or misleading, that there is fair balance of benefit/risk information, and that it is accurately communicated [62, 63].

Since 1962, the FDA was granted the responsibility to regulate prescription drug advertising and labeling [64, 65]. The regulations include reviewing written, printed, or graphic material accompanying a regulated product ("promotional labeling") and materials published in journals and newspapers, broadcast, and telephone communications systems [64, 66]. However, the FDA does not have the authority to require sponsors to submit promotional materials for approval prior to their use [67]. According to the Food, Drug and Cosmetics Act, manufacturers in their advertisements should include a brief summary which truthfully communicates the product's indication, major side effects and contraindications, major warnings, significant precautions, drug interactions, and they should present an adequate balance of risks and benefits. For broadcast ads, two options are available to communicate drug information: a brief summary or a toll-free telephone number or website [68].

Because manufacturers are not required to submit copies of advertisements at the time of initial dissemination nor copies of advertising at the time of initial publication, the FDA sees promotional materials only after they have been released or broadcasted [69, 70]. However, many manufacturers do submit their materials before airing to avoid future problems. Once an advertisement is disseminated, if it contains violative messages, the FDA can require corrective actions by means of untitled letters, warning letters, injunctions and consent decrees, referrals for criminal investigation, or prosecution and seizures [70].

Untitled letters or notices of violation are issued for less serious violations and they usually require the sponsor to discontinue use of false or misleading advertising materials. Warning letters are usually issued when there are more serious violations (e.g. repetitive misconduct or there is a potential for serious health risks to the public) [63]. Warning letters contain a statement that failure to respond may result in another regulatory action and that the FDA can initiate court proceedings for a seizure, injunction, or criminal prosecution [65]. Therefore, when manufacturers receive a warning letter, they are supposed to correct the problem immediately and disseminate the correct message using mailings and journals. However, a previous study showed that the FDA enforcement actions against false and misleading drug ads declined in 2002 and that there were delays in enforcement actions [71–73].

In November 2005, The Pharmaceutical Research and Manufacturers of America (PhRMA) issued some principles on the advertising of prescription drugs but the effect of those guidelines on warning letters is unknown. As a result of the above, Salas et al. described the number, type, and content of warning letters for prescribed medications and to assess if PhRMA guidelines had an effect on the number and content of warning letters issued [74]. They found that 25 % of the overall warning letters issued by the FDA were related directly with drugs and that 10 % were focused on drug-related promotional activities. They also found that half of the warning letters were issued because of superiority claims which encourage prescriber's not only to use of drugs but also to try the use of drugs for non- approved indications (i.e. off-label uses). In addition, they found an increase in warning letters issued in 1998 compared to previous years, which may be an effect of changes in the 1997 law. According to this law, the Food and Drug Administration Modernization Act of 1997 reauthorizes the Prescription Drug User Fee Act of 1992, regulating advertising of unapproved uses of approved drugs, and it released a draft guidance for direct to consumer advertising, which might have influenced an increase in the production of promotional materials [75].

Off-Label Drug Use (OLDU)

Off-label drug use is a commonly used term that many consider pejorative. However, OLDU is defined as available (marketed) medications that are prescribed for indications that have not received FDA approval (for a disease, symptom, and/or at a specific dose or dosage form—i.e. see definition above of "new drug" according to the FDA). OLDU can come about by virtue of the fact that the drug had not been studied in a specific population, or it is within a class of drugs that had already been approved for that specific population. Radley et al. reported that OLDU could constitute 21 % of all prescriptions [76, 77]. There are a number of questions regarding OLDU including whether a drug that is off-label can become standard of care, why, if the OLDU is beneficial FDA approval is not obtained, the legal vulnerability of using an off-label drug, whether drug companies can promote OLDU, and whether speakers can discuss them, and the difference between OLDU and orphan use of drugs. The fact is that off-label use of drugs can be commonly used and can become standard of care as exemplified by aspirin, which is approved for pain and fever, but not for the many coronary disease uses for which it is standard of care. As to why companies do not seek FDA approval for OLDU, obtaining such approval is costly and time-consuming, and may not be cost-effective particularly given that the medication is already being used for an off-label condition. An extensive discussion of the legal ramification of OLDU (should an adverse effect occur) is provided by Wittich et al., but they list four questions physicians should ask themselves when prescribing medications for OLDU as follows: "does the native drug have FDA approval", has the off-label use been subjected to substantial peer-review, is the off-label use medically necessary for treatment, and is the use of the medication non-experimental. Finally, speakers are allowed to discuss OLDU during their presentations as long as the drug's use is based on "*evidence that it is accepted within the profession of medicine*" [78]. But, pharmaceutical manufacturers are not allowed to promote off-label drug uses although they are "allowed to respond to unsolicited question from health care professionals about" OLDU; and, since 2009 are able to distribute journal articles and text-book chapters describing unapproved uses.

In summary, the USFDA has a long history of regulating new drug development, and in trying to insure the safety of drugs both before and after they reach the marketplace. The regulatory authority granted to the FDA is a dynamic process and the constant changes require continual updating of ones, knowledge.

References

1. Lewis S, Baird P, Evans RG, Ghali WA, Wright CJ, Gibson E, et al. Dancing with the porcupine: rules for governing the university-industry relationship. Can Med Assoc J. 2001; 165:783–5.
2. Kurian RG, editor. The historical guide to American government. New York: Oxford University Press; 1998.
3. Swann R. History of the FDA. Available from: www.fda.gov/oc/history. Cited 5 Sept 2007.
4. Guidance for Industry. Available from: www.fda.gov/cber/guidelines
5. Thelithromycin. Wikipedia. Available from: www.en.wikipedia.org/wiki/Telithromycin
6. Brott TG, Hobson RW, Howard G, Roubin GS, Clark WM, Brooks W, et al. Stenting versus endarterectomy for treatment of carotid-artery stenosis. N Engl J Med. 2010;363:11–23. PMC2932446.
7. Compton 3rd WM, Cottler LB, Jacobs JL, Ben-Abdallah A, Spitznagel EL. The role of psychiatric disorders in predicting drug dependence treatment outcomes. Am J Psychiatry. 2003; 160:890–5.
8. FDA Amendment Act of 2007. Available from: www.fda.gov/oc/initatives/hr3580.pdf
9. The Mission Statement of the ICH. Available from: http://www.ich.org
10. Coronary Drug Project. Available from: www.fda.gov
11. Kummar S, Rubinstein L, Kinders R, Parchment RE, Gutierrez ME, Murgo AJ, et al. Phase 0 clinical trials: conceptions and misconceptions. Cancer J. 2008;14:133–7. doi:10.1097/PPO.0b013e318172d6f3.
12. Suntharalingam G, Perry MR, Ward S, Brett SJ, Castello-Cortes A, Brunner MD, et al. Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. N Engl J Med. 2006;355:1018–28.
13. European Medicines Agency (EMEA). Available from: www.emea.europa
14. O'Donnell P. Not yet the last word on first-in-man. Appl Clin Trials. 2007;8:34–8.
15. Sullivan JT, McCarthy K. Decisions in first-in-human trials. Appl Clin Trials. 2010;2:58.
16. Palesch YY, Tilley BC, Sackett DL, Johnston KC, Woolson R. Applying a phase II futility study design to therapeutic stroke trials. Stroke. 2005;36:2410–4.
17. McCulloch P. Developing appropriate methodology for the study of surgical techniques. J Royal Soc Med. 2009;102:51–5.
18. Henderson L. The long arm of the FDA. Applied Clinical Trials, 2007 July.
19. DeMaria AN. The exportation of clinical research. J Am Coll Cardiol. 2009;53:1919–20. doi:10.1016/j.jacc.2009.04.019.
20. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ. 2003;326:1167–70.
21. Perlis RH, Perlis CS, Wu Y, Hwang C, Joseph M, Nierenberg AA. Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. Am J Psychiatry. 2005;162:1957–60.
22. Ridker PM, Torres J. Reported outcomes in major cardiovascular clinical trials funded by for-profit and not-for-profit organizations: 2000–2005. JAMA. 2006;295:2270–4.
23. Steri J. Point/Counterpoint Editorial. Cardiology News. Feb 2009
24. Kaiser KA, Cofield SS, Fontaine KR, Glasser SP, Thabane L, Chu R, et al. Is funding source related to study reporting quality in obesity or nutrition randomized control trials in top-tier medical journals? Int J Obes. 2012;36:977–81. PMC3288675.
25. Lee K, Bacchetti P, Sim I. Publication of clinical trials supporting successful new drug applications: a literature analysis. PLoS Med. 2008;5:1348–56. PMC2553819.
26. Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. BMJ. 2012;344:d7292. PMC3623605.
27. Zuckerman DM, Brown P, Nissen SE. Medical device recalls and the FDA approval process. Arch Intern Med. 2011;171:1006–11. doi:10.1001/archinternmed.2011.30.
28. Dhruva SS, Bero LA, Redberg RF. Strength of study evidence examined by the FDA in premarket approval of cardiovascular devices. JAMA. 2009;302:2679–85. doi:10.1001/jama.2009.1899.

29. Dhruva SS, Bero LA, Redberg RF. Gender bias in studies for Food and Drug Administration premarket approval of cardiovascular devices. Circ Cardiovasc Qual Outcomes. 2011;4:165–71. doi:10.1161/CIRCOUTCOMES.110.958215.
30. Institute of Medicine. Women's health research: progress, pitfalls, and promise. Washington, DC: National Academies Press; 2010.
31. Maisel WH. Medical device regulation: an introduction for the practicing physician. Ann Intern Med. 2004;140:296–302.
32. Kaplan AV, Baim DS, Smith JJ, Feigal DA, Simons M, Jefferys D, et al. Medical device development: from prototype to regulatory approval. Circulation. 2004;109:3068–72.
33. Ashton CM, Wray NP. Comparative effectiveness research: evidence, medicine, and policy. New York: Oxford University Press; 2013. p. 80–96.
34. Ramsey SD, Sullivan SD. Evidence, economics, and emphysema: Medicare's long journey with lung volume reduction surgery. Health Aff (Millwood). 2005;24:55–66. doi:10.1377/hlthaff.24.1.55.
35. Brantigan OC, Mueller E. Surgical treatment of pulmonary emphysema. Am Surg. 1957;23:789–804.
36. Cooper JD, Trulock EP, Triantafillou AN, Patterson GA, Pohl MS, Deloney PA, et al. Bilateral pneumectomy (volume reduction) for chronic obstructive pulmonary disease. J Thorac Cardiovasc Surg. 1995;109:106–16; discussion 16–9.
37. Tunis SR, Pearson SD. Coverage options for promising technologies: Medicare's 'coverage with evidence development'. Health Aff. 2006;25:1218–30.
38. Fishman A, Martinez F, Naunheim K, Piantadosi S, Wise R, Ries A, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. N Engl J Med. 2003;348:2059–73.
39. Kolata G. Medicare says it will pay, but patients say 'no thanks'. The New York Times on the Web. 2006:C1, C4. Available from: www.nytimes.com
40. Cooley DA, Al-Naaman YD, Carton CA. Surgical treatment of arteriosclerotic occlusion of common carotid artery. J Neurosurg. 1956;13:500–6.
41. Winslow CM, Solomon DH, Chassin MR, Kosecoff J, Merrick NJ, Brook RH. The appropriateness of carotid endarterectomy. N Engl J Med. 1988;318:721–7.
42. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Engl J Med. 1991;325:445–53.
43. Endarterectomy for asymptomatic carotid artery stenosis. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. JAMA. 1995;273:1421–8.
44. Barnett HJ, Taylor DW, Eliasziw M, Fox AJ, Ferguson GG, Haynes RB, et al. Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. N Engl J Med. 1998;339:1415–25.
45. Halm EA, Tuhrim S, Wang JJ, Rojas M, Hannan EL, Chassin MR. Has evidence changed practice?: appropriateness of carotid endarterectomy after the clinical trials. Neurology. 2007;68:187–94.
46. Blackshear JL, Cutlip DE, Roubin GS, Hill MD, Leimgruber PP, Begg RJ, et al. Myocardial infarction after carotid stenting and endarterectomy: results from the carotid revascularization endarterectomy versus stenting trial. Circulation. 2011;123:2571–8. PMC3173718.
47. Moseley JB, O'Malley K, Petersen NJ, Menke TJ, Brody BA, Kuykendall DH, et al. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med. 2002;347:81–8.
48. Kirkley A, Birmingham RB, Litchfield RB, Giffin JR, Willits KR, Wong CJ, et al. A randomized trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med. 2009;361:1097–107. doi:10.1056/NEJMoa0708333.
49. Katz JN, Brophy RH, Chaisson CE, de Chaves L, Cole BJ, Dahm DL, et al. Surgery versus physical therapy for a meniscal tear and osteoarthritis. N Engl J Med. 2013;368:1675–84. PMC3690119.

50. Wenner DM, Brody BA, Jarman AF, Kolman JM, Wray NP, Ashton CM. Do surgical trials meet the scientific standards for clinical trials? J Am Coll Surg. 2012;215:722–30. PMC3478478.
51. Weinstein JN, Tosteson TD, Lurie JD, Tosteson AN, Hanscom B, Skinner JS, et al. Surgical vs. nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. JAMA. 2006;296:2441–50.
52. Brody BA. The ethics of biomedical research: an international perspective. New York: Oxford University Press; 1998.
53. Kim SY, Frank S, Holloway R, Zimmerman C, Wilson R, Kieburtz K. Science and ethics of sham surgery: a survey of Parkinson disease clinical researchers. Arch Neurol. 2005;62(9):1357–60.
54. Wennberg JE. An apple a day? N Engl J Med. 1994;331:815; author reply 6.
55. The AMIS Study Group. The aspirin myocardial infarction study: final results. The Aspirin Myocardial Infarction Study research group. Circulation. 1980;62:79–84.
56. Harlow SD, Linet MS. Agreement between questionnaire data and medical records. The evidence for accuracy of recall. Am J Epidemiol. 1989;129:233–48.
57. St Sauver JL, Hagen PT, Cha SS, Bagniewski SM, Mandrekar JN, Curoe AM, et al. Agreement between patient reports of cardiovascular disease and patient medical records. Mayo Clin Proc. 2008;80:203–10. http://dx.doi.org/10.4065/80.2.203.
58. Bolland MJ, Barber A, Doughty RN, Grey A, Gamble G, Reid IR. Differences between self-reported and verified adverse cardiovascular events in a randomised clinical trial. BMJ Open. 2013;3:1–6. PMC3612743.
59. The FDA's poor oversight of postmarketing studies. (Editorial). The Lancet. 2009;374:1568. doi: 10.1016/S0140-6736(09)61932-2.
60. Strain EC, Stitzer ML, Bigelow GE. Early treatment time course of depressive symptoms in opiate addicts. J Nerv Ment Dis. 1991;179:215–21.
61. Hesse M. The Beck Depression Inventory in patients undergoing opiate agonist maintenance treatment. Br J Clin Psychol. 2006;45:417–25.
62. 21 CFR Part 310 section 502(a) of the Food and Drug Administration Modernization Act of 1997, 21 U.S.C. 352(a).
63. Baylor-Henry M, Drezin N. Regulation of prescription drug promotion: direct-to consumer advertising. Clin Ther. 1998;20(C):C86–95.
64. Section 502(n) of the Food Drug and Cosmetics Act, and Title 21 Code of Federal Regulations, 202.1(1) (1).
65. Kessler DA, Pines WL. The federal regulation of prescription drug advertising and promotion. JAMA. 1990;264:2409–15.
66. 21 CFR Part 310 section 502(a) of the Food and Drug Administration Modernization Act of 1997 (Modernization Act)., 21 U.S.C. 352 (n).
67. Rounsaville BJ, Weissman MM, Crits-Christoph K, Wilber C, Kleber H. Diagnosis and symptoms of depression in opiate addicts. Course and relationship to treatment outcome. Arch Gen Psychiatry. 1982;39:151–6.
68. Section 502 (n) of the Food Drug and Cosmetics Act, and Title 21 Code of Federal Regulations, 202.1.
69. Chen LY, Crum RM, Martins SS, Kaufmann CN, Strain EC, Mojtabai R. Service use and barriers to mental health care among adults with major depression and comorbid substance dependence. Psychiatr Serv. 2013;64:863–70. doi:10.1176/appi.ps.201200289.
70. Woodcock J. Statement by Janet Woodcock, MSD Director, Center of Drug Evaluation and Research. US Drug Administration. Department of Health and Human Services; 2003.
71. Gahart MT, Duhamel LM, Dievler A, Price R. Examining the FDA's oversight of direct to consumer advertising. Health Aff (Millwood). 2003;Suppl Web Exclusives:W3–120–3. doi: 10.1377/hlthaff.w3.120.
72. Waxman HA. Ensuring that consumers receive appropriate information from drug ads: What is the FDA's role? Health Aff (Millwood). 2004;Suppl Web Exclusives:W4–256–8. doi: 10.1377/hlthaff.w4.256.

73. Waxman RHA. Letter from Rep. Henry A. Waxman to the Honorable Tommy G. Thompson. 2002. Available from: www.oversight.house.gov/story.asp?ID=441. Cited 1 June 2007.
74. Salas M, Martin M, Pisu M, McCall E, Zuluaga A, Glasser SP. Analysis of US food and drug administration warning letters: false promotional claims relating to prescription and over-the-counter medications. Pharm Med. 2008;22:119–25.
75. Food and Drug Administration. Available from: www.fda.gov/opacom/backgrounders/miles.html. Cited 10 June 2010.
76. Radley DC, Finkelstein SN, Stafford RS. Off-label prescribing among office-based physicians. Arch Intern Med. 2006;166:1021–6.
77. Wittich CM, Burkle CM, Lanier WL. Ten common questions (and their answers) about off-label drug use. Mayo Clin Proc. 2012;87:982–90. PMC3538391.
78. Accreditation Council for Continuing Medical Education Website. Available from: www.accme.org/education. Accessed 17 Feb 2013.

Chapter 7
The Placebo and Nocebo Effect

Stephen P. Glasser and William Frishman

> *If a placebo were submitted to the FDA for approval, they would no doubt be impressed with its efficacy, but would probably not approve it due to its frequent side effects.*

Abstract There are four general reasons for clinical improvement in a patient's condition: (1) natural history of the disease; (2) specific effects of the treatment; (3) regression to the mean; and (4) nonspecific effects of the treatment that are attributable to factors other than the specific active components. The latter effect is included under the heading 'placebo effect'. In this chapter the placebo effect will be discussed, with some emphasis on regression to the mean. Placebos ('I will please') and their lesser known counterpart's nocebo's (I will harm') are sham treatments. The difference is in the response to the inert therapy. A beneficial response to an inert substance is a placebo response; a side effect to an inert substance is a nocebo response.

Keywords Placebo • Nocebo • Regression to the mean • Placebo mechanisms • Placebo in clinical trials • Placebo ethics • Placebo characteristics

Placebo has been cited in PubMed over 170,000 times indicating that placebo has set the standard for how clinical research and particularly clinical trials are conducted. On the other hand, some have argued that placebo effects are overstated and can be explained by other variables (e.g. changes in the natural history of the disease, regression to the mean, methodological issues, conditioned answers, etc.). The importance,

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

W. Frishman, M.D., MACP
The Department of Medicine, New York Medical College, New York City, NY, USA

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_7,

Table 7.1 Four general reasons for clinical improvement in a patient's condition

Natural history of the disease
Specific effects of the treatment
Regression to the mean
Placebo effect

controversy, and to date inadequate study of the placebo effect, warrants an in depth review of this topic. In addition, the discussion of placebos requires an understanding of the ethics of clinical trials, intention to treat analysis, surrogate endpoints and many of the other areas that have already been discussed in prior chapters.

Placebos ('I will please') and their lesser-known counterpart's nocebos ('I will harm') are sham treatments. The difference between placebo and nocebo is in the response to the inert therapy. A beneficial response to an inert substance is a placebo response; a side effect to an inert substance is a nocebo response.

There are four general reasons for clinical improvement in a patient's condition: (1) natural history of the disease; (2) specific effects of the treatment; (3) regression to the mean; and (4) nonspecific effects of the treatment that are attributable to factors other than the specific active components (Table 7.1). The latter effect is included under the heading 'placebo effect' [1]. Each time a physician recommends a diagnostic or therapeutic intervention for a patient, built into this clinical decision is the possibility of a placebo effect, that is, a clinical effect unrelated to the intervention itself [2]. Simple diagnostic procedures such as phlebotomy or more invasive procedures such as cardiac catheterization have been shown to have important associated placebo effects [3, 4]. Chalmers [5] has stated that a simple review of the many abandoned therapies reveals that many patients would have benefited by being assigned to a placebo control group. In fact, what might represent the first known clinical trial, and one in which the absence of a placebo control group led to erroneous conclusions, is a summary attributed to Galen in 250 BC, who stated that 'some patients that have taken this herbivore have recovered, while some have died; thus, it is obvious that this medicament fails only in incurable diseases' [6].

Placebo effects are commonly observed in patients with cardiac disease who also receive drug and surgical therapies as treatments. Rana et al. noted the 'tremendous power of the placebo effect' in patients with end-stage coronary disease in clinical trials of angiogenesis and laser myocardial revascularization [7]. They also commented on the fact that the observed improvements were not limited to 'soft' symptomatic endpoints but were also observed with 'hard' endpoints such as exercise walking time on a treadmill, and in magnetic resonance imaging. Rana et al. also studied the longevity of the placebo effect from published clinical trials. They found that the beneficial effects of placebo (on angina class, angina frequency, and exercise time) persisted for up to 2 years.

Definition

Stedman's Medical Dictionary [7] defines the word 'placebo,' which originates from Latin verb meaning 'I shall please,' to have two meanings. First, a placebo may be an inert substance prescribed for its suggestive value. Second, it may be an inert substance identical in appearance with the compound being tested in experimental research, and the use of which may or may not be known by the physician or the patient; it is given to distinguish between the action of the compound and the suggestive effect of the compound under study [8].

Currently, there is some disagreement as to the exact definition of a placebo. Many articles on the subject include a broader definition, as given by Shapiro in 1961 [9].

> "*Any therapeutic procedure (or that component of any therapeutic procedure) which is given deliberately to have an effect or unknowingly has an effect on a patient, symptom, syndrome, or disease, but which is objectively without specific activity for the condition being treated. The therapeutic procedure may be given with or without conscious knowledge that the procedure is a placebo, may be an active (noninert) or nonactive (inert) procedure, and includes, therefore, all medical procedures no matter how specific—oral and parenteral medication, topical preparations, inhalants, and mechanical, surgical and psychotherapeutic procedures. The placebo must be differentiated from the placebo effect, which may or may not occur and which may be favorable or unfavorable. The placebo effect is defined as the changes produced by placebos. The placebo is also used to describe an adequate control in research.*"

A further refinement of the definition was proposed by Byerly [10] in 1976 as '*any change in a patient's symptoms that are the result of the therapeutic intent and not the specific physiochemical nature of a medical procedure*.'

Placebo Effect in Clinical Trials

The use of placebo controls in medical research was advocated in 1753 by Lind [11] in an evaluation of the effects of lime juice on scurvy. After World War II, research protocols designed to assess the efficacy and safety of new pharmacologic therapies began to include the recognition of the placebo effect.

The roots of the placebo problem can be traced to a lie told by an Army nurse during World War II as Allied forces stormed the beaches of southern Italy. The nurse was assisting an anesthetist named Henry Beecher, who was tending to US troops under heavy German bombardment. When the morphine supply ran low, the nurse assured a wounded soldier that he was getting a shot of potent painkiller, though her syringe contained only salt water. Amazingly, the bogus injection relieved the soldier's agony and prevented the onset of shock.

Returning to his post at Harvard after the war, Beecher became one of the nation's leading medical reformers. Inspired by the nurse's healing act of deception, he launched a crusade to promote a method of testing new medicines to find out whether they were truly effective. At the time, the process for vetting drugs was sloppy at best, and Pharmaceutical companies would simply dose volunteers with an experimental agent until the side effects swamped the presumed benefits. Beecher proposed that if test subjects could be compared to a group that received a placebo, health officials would finally have an impartial way to determine whether a medicine was actually responsible for making a patient better.

Placebos and their role in controlled clinical trials were recognized in 1946, when the Cornell Conference on Therapy devoted a session to placebos and double-blind methodology. At that time, placebos were associated with increased heart rate, altered respiration patterns, dilated pupils, and increased blood pressure. In 1951, Hill [12] concluded that a change in a patient to be attributable to a specific treatment (for better or worse) the result must be repeatable a significant number of times in other similar patients. Otherwise, the result could be due simply to the natural history of the disease or the passage of time. He also proposed the inclusion of a control group that received identical treatment except for the exclusion of an 'active ingredient.' Thus, the 'active ingredient' was separated from the situation within which it was used. This control group, also known as a placebo group, would help in the investigations of new and promising pharmacologic therapies.

Beecher [13] was among the first investigators to promote the inclusion of placebo controls in clinical trials. He emphasized that neither the subject nor the physician should know what treatment the subject was receiving and referred to this strategy as the 'double unknown technique.' Today, this technique is called the 'double-blind technique' and ensures that the expectations and beliefs of the patient and physician are excluded from evaluation of new therapies. In 1955, Beecher reviewed 15 studies that included 1,082 patients and found that an average of 35 % of these patients significantly benefited from placebo therapy (another third had a lesser benefit). He also concluded that placebos can relieve pain from conditions with physiologic or psychological etiologies. He described diverse objective changes with placebo therapy. Some medical conditions improved; they included severe postoperative wound pain, cough, drug-induced mood changes, pain from angina pectoris, headache, seasickness, anxiety, tension, and the common cold.

The Use of Placebos in Clinical Trials

There has been renewed interest in the use of placebos in clinical trials, and, not just because of the ethical issues involved. For example, from 2001 to 2006, the percentage of new products dropped from development after Phase II clinical trials, when drugs are generally first tested against placebo, rose by 20 %. During that same time period the failure rate in more extensive Phase III trials increased by 11 %, mainly as the result of surprisingly poor showings against placebo. Also, half of all drugs that fail in late-stage trials drop out of the pipeline due to their inability to beat placebo. Some examples are: a new type of gene therapy for Parkinson's disease was abruptly withdrawn from Phase II trials after unexpectedly tanking against placebo, stem-cell trials for Crohn's disease were suspended citing an "unusually high" response to placebo, and clinical trials for a much-touted new drug for schizophrenia was stopped when volunteers showed double the expected level of placebo response. And, it's not only trials of new drugs that are crossing the futility boundary. Some products that have been on the market for decades are faltering in more recent follow-up tests, and in many cases, these are the compounds that, in the late 1990s, made Big Pharma more profitable than Big Oil, yet if these same drugs were studied now, the FDA might not approve some of them. Further confounding things is the observation that while some drugs are more likely to be superior in American studies than in those done in Europe and South Africa, others are still beating placebo in France and Belgium, but not in the USA.

Finally, since the 1980s, two comprehensive analyses of antidepressant trials have uncovered a dramatic increase in placebo response. One estimated that the effect size in placebo groups had nearly doubled over that time; and, it's not that the old treatments are getting weaker, it's as if the placebo effect is somehow getting stronger.

Characteristics of the Placebo Effect

There appears to be an inverse relation between the number of placebo doses that needs to be administered and treatment outcomes. In a study of patients with postoperative wound pain, 53 % of the subjects responded to one placebo dose, 40 % to two or three doses, and 15 % to four doses [12]. In analyzing the demographics of those who responded to placebo and those who did not, Lasagna et al. [14] found no differences in gender ratios or intelligence quotients between the two groups. They did find significant differences in attitudes, habits, educational backgrounds, and personality structure between consistent responders and nonresponders. In attempting to understand the reproducibility of the placebo effect, some have observed that there was no relation between an initial placebo response and subsequent responses with repeated placebo doses of saline [12]. Beecher concluded that placebos are most effective when stress, such as anxiety and pain, is greatest. But, placebo responses can be associated with dose response characteristics, frequency of dosing, pill color (e.g. blue vs. pink pills are more sedating, yellow vs. green more stimulating) and, "branded placebo" in some studies were more effective than generic placebo (Fig. 7.1). The magnitude of effect is difficult to quantitate due to its diverse nature but it is estimated that a placebo effect accounts for 30–40 % of an interventions benefit.

Placebos can produce both desirable and adverse reactions. Some now use the term placebo for the beneficial effects and nocebo for the adverse effects. Beecher et al. described >35 adverse reactions from placebos; the most common are listed in Table 7.2. The aforementioned reactions were recorded without the patient's or physician's knowledge that a placebo had been administered. In one study in which lactose tablets were given as a placebo, major adverse reactions occurred in three patients [15]. The first patient had overwhelming weakness, palpitation, and nausea after taking both the placebo and then the test drug. In the second patient, a diffuse rash developed with placebo administration, and the rash disappeared after placebo was discontinued. The third patient had epigastric pain followed by watery diarrhea, urticaria, and angioneurotic edema of the lips after receiving the placebo.

Yellow pills
make the most effective antidepressants

Red pills
can give you a more stimulating kick

The color green
reduces anxiety

More is better,
Placebos taken four times a day
deliver greater relief than those taken twice daily.

Placebos Are Getting More Effective. Drugmakers Are Desperate to Know Why.
Steve Silberman http://www.wired.com/medtech/drugs/magazine/17-09/ff_placebo_effect?currentPage=all

Fig. 7.1 Pill color and its placebo effects

Table 7.2 Common adverse reactions to Placebo (Nocebo effect)

Reaction	Incidence (%)
Drowsiness	50
Headache	25
Sensation of heaviness	18
Fatigue	18
Difficulty concentrating	15
Sleep disturbance	10
Nausea	10
Overly relaxed	9

Indeed, because of the substantial evidence of placebo ‘efficacy’ and placebo ‘side effects,’ some investigators have wittingly suggested that if placebo were submitted to the United States Food and Drug Administration (FDA) for approval, that the agency, though impressed with the efficacy data, would probably recommend disapproval on the basis of the high incidence of side effects. Some authors have questioned whether placebos are truly inert. Davis pointed out that part of the problem with the placebo paradox is our failure to separate the use of an inert medication (if there is such as substance) from the phenomenon referred to as the placebo effect. It might help us if we could rename the placebo effect the “obscure therapeutic effect” [16].

For instance, in trials of lactase deficiency therapy, could the amount of lactose in placebo tablets actually cause true side effects? Although the small amount of lactose makes this possibility seem unlikely. Perhaps it is more likely that allergies to some of the so-called inert ingredients in placebos cause reactions in predisposed persons, although this explanation probably could not explain more than a small percentage of placebo side effects.

A validation of the placebo effect occurred in 1962 when the United States enacted the Harris-Kefauver amendments to the Food, Drug, and Cosmetic Act. These amendments required proof of efficacy and documentation of relative safety, in terms of the risk-benefit ratio for the disease to be treated, before an experimental agent could be approved for general use [17]. In 1970, the FDA published rules for 'adequate and well-controlled clinical evaluations.' The federal regulations identified five types of controls (placebo, dose-comparison, active, historical, and no treatment) and identified use of the placebo control as an indispensable tool to achieve the standard [18]. However, the FDA does not mandate placebo controls, and in fact has stated that placebo groups are 'desirable, but need not be interpreted as a strict requirement. The speed with which blind comparisons with placebo and/or positive controls can be fruitfully undertaken varies with the nature of the compound. In the publication regarding 'Draft Guidelines for the Clinical Evaluation of Anti-anginal Drugs,' the FDA further states that '*it should be recognized that there are other methods of adequately controlling studies. In some studies, and in some diseases, the use of an active control drug rather than a placebo is desirable, primarily for ethical reasons.*"

Regression Towards the Mean (or Towards Mediocrity)

An important statistical concept and one that may mimic a placebo response or a clinical response is regression towards the mean or regression towards mediocrity (RTM). RTM identifies a phenomenon that a biologic variable that is extreme on its first measurement will tend to be closer to the center of the distribution on a later measurement. The term originated with Sir Francis Galton who studied the relationship between the height of parents and their adult offspring. He observed that children of tall parents were (on average) shorter than their parents; while, children of short parents were taller than their parents. Galton called this regression towards mediocrity [20]. Another example of RTM is from Ederer, who observed that during the first week of the 1968 baseball season the top ten and bottom ten batters averaged 0.414 and 0.83 respectively. The following week they hit 0.246 and 0.206 respectively, while the average for the league remained stable [19].

At least three types of studies are potentially affected by RTM: a survey in which subjects are selected for subsequent follow-up based upon an initial extreme value, studies with no control groups, and even controlled trials. An example is taken from the Lipid Research Clinics Prevalence Study, a sample population who had elevated total cholesterol was asked to return for reevaluation. According to RTM, it would be expected that the 2nd measurement would on average be lower, and this would not be so had a randomly selected sample been chosen for reevaluation [22].

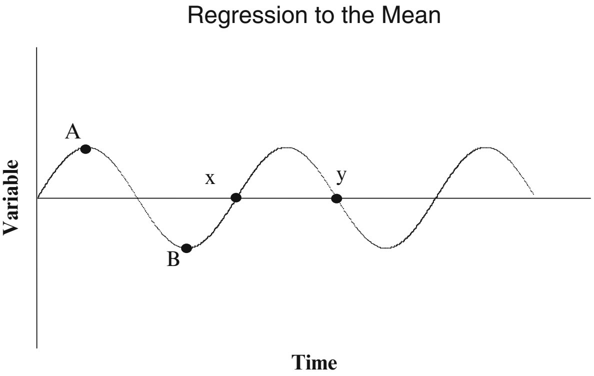


Fig. 7.2 If one measures a variable at its peak value (A in the example) the next measurement is likely to be lower (B, x, or y in this example). Conversely, if one were to measure a variable at its lowest point (B), the next measurement is likely to be higher

The reason that a randomly selected sample would be less likely to demonstrate RTM is because the random sample would have representative values across the spectrum of cholesterol measurements at the start, whereas the selected sample all initially had elevated values.

Another example of the RTM principal comes from the National Diet-Heart Study [23]. It had been repeatedly observed that a low cholesterol diet given to subjects with high cholesterol values resulted in greater cholesterol lowering that when the same diet was given to someone with lower cholesterol values. In the National Diet-Heart Study subjects with a baseline cholesterol >242 mg/dL had a 15 % reduction while those whose baseline cholesterol was 210–241 mg/dL had a 12 % reduction [23]. There are two possible explanations for this observation: one, that the diet hypothesis holds i.e. that subjects with high cholesterol are more responsive to cholesterol lowering treatment than those with lower cholesterol values; and two, that independent of dietary intervention subjects with high cholesterol will (on average) decrease more than those with lower values due to RTM. In fact, it is likely that both could occur simultaneously.

RTM then, is a phenomenon that can make a natural variation in repeated data look like a real change. In biologic systems, most variables increase and decrease around a mean (as, for instance, might be visualized as a sine wave). Thus, it is likely that any value measured at a specific point in time will, by chance, either be above or below the mean, and that a second measurement will be at a different point around the mean and, therefore, different from the first measurement (Fig. 7.2). The presumption is that this variability about the mean will be the same in the placebo group as in the active treatment group (assuming adequate sample size and randomization), so that differences between the two groups relative to regression to the mean will cancel out. In an intervention study, RTM cannot be observed because it is mixed into the genuine intervention effect. This is particularly true of intervention studies where the population selected for study generally is in the high risk groups—that is with values that are high at baseline. Yudkin and Stratton evaluated this by analyzing a group with high baseline cholesterol, and observing a 9 % fall without any intervention [21]. These authors go on to point out

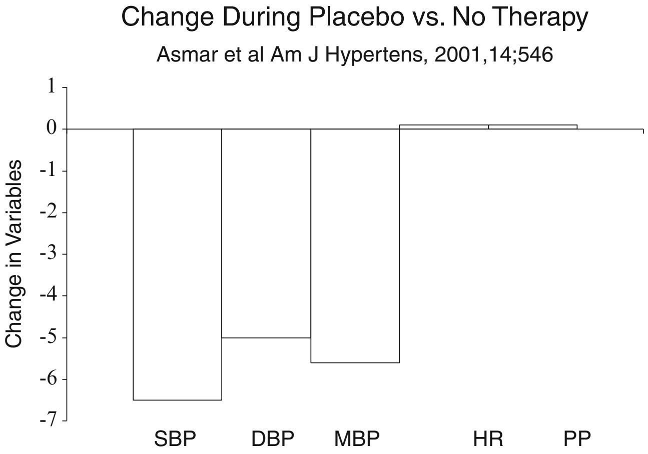


Fig. 7.3 Change in measured variables during placebo vs. no therapy. From Asmar et al. [22]

several ways of estimating the impact of RTM, and three suggested approaches to minimizing the RTM problem. These approaches include the use of an RCT design, since the RTM effect will be part of the total effect of the response in both the intervention and control groups. However, the response in both groups will be inflated by the RTM so the true impact of the intervention is not known and is likely somewhat less that that observed. A second approach to minimizing RTM is to obtain several measurements and average them to determine baseline. The third approach is to use the first measurement as the basis for selection of the subject into the study, and a second measurement that will be used as the baseline from which to assess the effect of the intervention.

The ideal comparator for a study would actually be no therapy vs. the investigational agent, however, the loss of blinding makes this approach problematic as well. There has been little study of the no therapy control, however, Asmar et al. did attempt to evaluate this as part of a larger interventional trial [22]. They used a randomized cross-over approach with a 1 month run-in followed by a 1 month placebo vs. no treatment period. BP and ABPM were measured. The results could be then analyzed in terms of the no treatment effect (no parameters changed in the two periods) and the RTM effect shown in Fig. 7.3.

Mechanism of the Placebo Effect

There has been much discussion regarding the mechanism of the placebo response. However, the mechanism at the cellular level and the role of biochemical mediators continues to escape detection. In an attempt to elucidate some mechanisms of the placebo effect, Beecher [13] described two phases of suffering: first, the initial pain sensation or other symptom, and second the person's reaction to this sensation or experience by the central nervous system. The first, or somatic, phase is associated with the source of the pain or symptom; the second, or cortical, phase is superimposed on the pain or symptom. An example of the influence of the effect of the mind on the body is the 'Anzio Effect.' During World War II, injured soldiers at Anzio, Italy, complained less of pain after surgery, than typical patients after surgery. This difference was recognized because less than one third of the injured soldiers required morphine, compared with four fifths of patients undergoing similar recovery from the same surgery in non-combatants. For the soldiers, the knowledge that they had survived, combined with the anticipation of returning home, probably reduced their pain. In contrast, typical surgical patients are required to comply with hospital procedures, probably producing anxiety or fear that acts to increase pain [23]. The physiologic mechanism involved with pain begins when fear or anxiety activates the hypothalamus-hypophysis-adrenal axis, resulting in release of catecholamines. These catecholamines act on the body, which then sends feedback to the cerebral cortex via neural connections. The thalamus in the diencephalons, which processes sensory input before relaying it to the cerebral cortex, then sends recurrent axons to the thalamus, presumably to allow modulation of the input received from the thalamus [23, 24].

One theory to explain the placebo effect is classical conditioning, the pairing of an unconditioned stimulus with a conditioned stimulus until eventually the conditioned stimulus alone elicits the same response as the unconditioned stimulus. This effect of the environment on behavior was tested in a study by Voudouris et al. [25]. They studied responses to pain stimulation with and without a placebo cream. A visual analogue scale determined pain perception. To evaluate the effect of verbal expectancy, the patients were informed that the placebo cream had powerful analgesic properties (expectancy) or that the cream was neutral (no expectancy). To determine the role of conditioning, the level of pain stimulus was reduced after application of the cream (conditioning) or was maintained at the same level of pain (no conditioning). The patients were divided into four groups: a group receiving expectancy and conditioning, a group receiving only expectancy, a group receiving only conditioning, and a group receiving neither. Both conditioning and verbal expectancy were important mediators on the placebo response, but conditioning was more powerful [25].

A second explanation for the placebo effect is response by neurohormones, including motor or autonomic nervous systems, hormone systems, and immune systems. Endogenous neuroendocrine polypeptides, including β-endorphins, enkephalins, and antiopioids, are activated by many factors. These factors include placebos, vigorous exercise, and other stressors. Modulation of the opioid system may occur by an antiopid system of neurotransmitters. γ-Aminobutyric acid, and peptide neurotransmitter, is associated with the secretion of β-endorphin and β-lipotropin [23]. The endorphin group of neurotransmitters is created from the proopiomelanocortitrophin peptide and is linked through β-lipotropin with the regulation of the hypothalamus-hypophysis-adrenal axis. There is no understanding of the exact link between the opioid-antiopioid and β-lipotropin systems of neuroendocrine peptides. The brain peptides and their actions on presynaptic and postsynaptic receptors on neurons also are not understood. Experiments in animals provide most of the information about control of the genetic expression of the peptides [23].

In a double-blind study by Levine et al. [26], patients received placebo and then intravenous naloxone after tooth extraction. Naloxone, a partial opioid antagonist that competes with β-endorphins for the same receptor in the brain, blocked the placebo effect previously experienced by the patients. Levine et al. concluded that placebo activates β-endorphins in the brain and that naloxone increases the pain by inhibiting the placebo effect [26]. A double-blind study by Hersh et al. found ibuprofen to be more efficacious than placebo or codeine [27]. Naltrexone, a long-acting oral form of naloxone, given before oral surgery reduced the analgesic response to placebo and to codeine received after surgery. In an additional noteworthy finding, pretreatment with naltrexone prolonged the duration of ibuprofen's action rather than diminishing the peak analgesic response. This prolongation of ibuprofen's action was hypothesized to result from increased central stimulation of endogenous opiates by ibuprofen or from competition by naltrexone for liver enzymes involved in the inactivation and elimination of ibuprofen.

A third model of the placebo response is the ability of mental imagery to produce specific and measurable physiologic effects. This model explains the relation between psychological and physiologic components of the placebo effect. There is a conversion in the brain of psychological placebo-related imagery into a physiologic placebo response. A patient may modify his or her imagery content in response to bodily reactions during treatment, in response to the behaviors and attitudes of doctors or nurses, or in response to information about the treatment from other sources (such as other patients, books, and journals) [28]. An example of this model is described in another study [29]. Two matched groups of patients preparing to undergo abdominal surgery received different types of care. In one group, the anesthesiologist told the patients about the operation but not about the postoperative pain. The other group was told about the postoperative pain and assured that medication was available. It was found that the patients informed about postoperative pain needed only half the analgesic and left the hospital 2 days earlier. The authors concluded that this result showed 'a placebo effect without the placebo' [29].

Additional studies have been attempted to both characterize and explore the mechanisms of the placebo effect. One such approach has been based upon the color and shape of pills and how that affects how patients feel about their medication. For example, *ScienceDaily (Jan. 19, 2011)* reported that according to recent research the color, shape, taste and even name of a tablet or pill may have an effect on how patients feel about their medication. Choose an appropriate combination and the placebo effect gives the pill a boost, improves outcomes and might even reduce side effects. In fact, it has been observed that pill color may influence both the placebo and the nocebo effects (Fig. 7.1). Some general observations from this line of research suggests that capsules tend to be more effective than other pill forms, and that red and pink tables are generally more effective than other colors. A study was performed in order to assess the impact of the color of a drug's formulation on its perceived effect and its effectiveness, and to examine whether antidepressant drugs available in the Netherlands are different in color from hypnotic, sedative, and anxiolytic drugs [33]. The systematic review was of 12 published studies of which six examined the perceived action of different colored drugs and six the influence of the color of a drug on its effectiveness. The studies on perceived action of drugs showed that red, yellow, and orange were associated with a stimulant effect, while blue and green were related to a tranquillizing effect. The analysis of the studies that assessed the impact of the color of drugs on their effectiveness showed inconsistent differences between colors. However, hypnotic, sedative, and anxiolytic drugs were more likely than antidepressants to be green, blue, or purple. Their overall conclusions were that colors affect the perceived action of a drug and may influence the effectiveness of some drugs, that a relation exists between the coloring of drugs that affect the central nervous system and the indications for which they are used, and that further research contributing to a better understanding of the effect of the color of drugs is warranted [33].

Placebo Effect in Various Diseases

Placebo Effect in Ischemic Heart Disease and Chronic, Stable, Exertional Angina Pectoris

The rate of improvement in the frequency of symptoms in patients with chronic, stable, exertional angina pectoris with placebo therapy has been assessed to be 30–80 % [30]. A summary of subjective and objective placebo effects in cardiovascular disease is provided in Table 7.3. Because of the magnitude of the placebo effect, most studies of new antianginal therapies were performed with placebo control. However, the safety of this practice came under scrutiny in the late 1980s because of concern that patients with coronary artery disease would have periods of no drug treatment. As a result, Glasser et al. explored the safety of exposing patients with chronic, stable, exertional angina to placebos during short-term drug trials with an average double-blind period of 10 weeks [31]. The study included all new drug applications (NDAs) submitted to the FDA between 1973 and 2001. The results of these drug trials were submitted, whether favorable or not, and all adverse events were reported. Qualifying studies used symptom-limited exercise tolerance testing as an end point. No antianginal medication, except sublingual nitroglycerin, was taken after a placebo-free or drug-free washout period. A total of 2,921 patients with angina pectoris and an abnormal exercise tolerance test who entered any randomized, double-blind, placebo-controlled trial. Since then, an additional 9 NDAs (representing 63 trials) for angina claims have been submitted to the FDA, resulting in an updated total of 10,865 patients, among whom 607 (5.6 %) were withdrawn from the trials due to an adverse drug event. The relative risk (RR) for withdrawal (placebo compared to drug-treated patients) was not increased ($RR=0.92$, 0.78, 1.08; $p=0.28$).

Table 7.3 Objective placebo effects in cardiovascular disease

	Placebo effect
Heart failure [37]	
Exercise tolerance testing	
1 or 2 baseline measurements	90–120 s
3–10 baseline measurements	10–30 s
Increase in ejection fraction of 5 %	20–30 % of patients
Hypertension [53]	
Measured by noninvasive automatic ambulatory 24-h monitoring	0 %
Arrhythmia	
Study 1 [63][a]	
A reduction in mean hourly frequency of ventricular tachycardia	<65 %
A reduction in mean hourly frequency of couplets	<75 %
A reduction in mean hourly frequency of all ventricular ectopic beats without regard for complexity	<83 %
Study2 [64][b]	
Baseline VPCs > 100/h	<3 times baseline
Baseline VPCs < 100/h	<10 times baseline
Silent ischemic disease [24]	
Reduction in frequency of ischemic events	44 %
Reduction in ST-segment integral	50 %
Reduction in duration of ST-segment depression	50 %
Reduction of total peak ST-segment depression	7 %
Other [67, 69, 72]	
Compliance with treatment at rate of ≥75 %	<3 times baseline

VPC Ventricular premature complexes

[a] Based on comparison of one control 24 h monitoring period to one 24-h treatment period. Variability is so great that it may be inadvisable to pool individual patient data to detect trends in ectopic frequency in evaluating new potential antiarrhythmic agents in groups of patients

[b] When differentiating proarrhythmia in patients with mixed cardiac disease and chronic ventricular arrhythmias from spontaneous variability, with false-positive rate of only 1 %

Combined events, irreversible harm (CVA, MI, Death), and serious cardiovascular events (MI, CHF, CVA) also had point estimates favoring randomization to placebo (RR = 0.54, 0.26, 1.04; $p < 0.068$ and RR = 0.89; .61, 1.30; $p = 0.56$ respectively). The conclusion was that with a greater number of trials and larger numbers of randomized patients, the results are similar to those reported prior; and, within the limitations of the study, there was no evidence that the use of a placebo control is unsafe in short-term studies of chronic stable angina (Fig. 7.4). This analysis found evidence that supported the safety of a placebo group in short-term drug trials for chronic, stable, exertional angina [37]. An analysis of the safety of a placebo control in trials of anti-hypertensive drugs has also been published [38]. Although a slightly increased risk of reversible symptoms was identified, there was no evidence of irreversible harm as a result of participation in any of these trials. The same caveats apply as discussed in the angina trials-that is, these were short term trials of carefully monitored and selected patients.

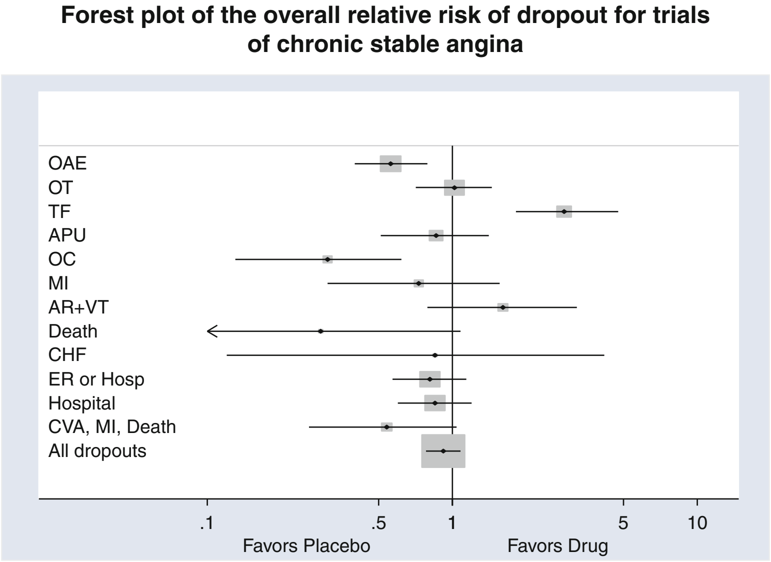


Fig. 7.4 Forest plot of the overall relative risk of dropout for trials of chronic stable angina. From: Glasser et al. [81]

The safety of using placebo in longer-term drug trials for chronic, stable, exertional angina has not been established. A placebo-controlled trial by a European group in 1986 enrolled 35 patients and made observations during a 6-month period of placebo or short-acting nitroglycerin administration [32]. This study of the long-term effects of placebo treatment in patients with moderately severe, stable angina pectoris found a shift toward the highest dosage during the titration period. Seven patients continued to receive the lowest dosage, but the average ending dosage was 65 % more than the initial dosage. Compliance, when determined by pill count, for 27 patients was >80 %. During the first 2.5 months of the trial, noncompliance with the regimen or physical inability to continue to study was ascertained. No patients died or had myocardial infarction [32].

There is a paucity of information regarding any gender differences in placebo response. Women represented 43 % of the population in the aforementioned European study [32] and were more likely to have angina despite normal coronary arteries. Because the placebo effect may be more pronounced in patients with normal coronary arteries, data from men were analyzed separately to compare them with the overall results. However, the data from men were very similar to the overall results. In fact, the functional status of men showed more improvement attributable to placebo (61 %) than overall (48 %) at 8 weeks. The results of this study showed no adverse effects of long-term placebo therapy: 65 % of patients reported subjective, clinical improvement and 27 % of patients reported objective, clinical improvement in exercise performance [32]. Of note, improvement in exercise performance can occur when patients undergo repeated testing [33].

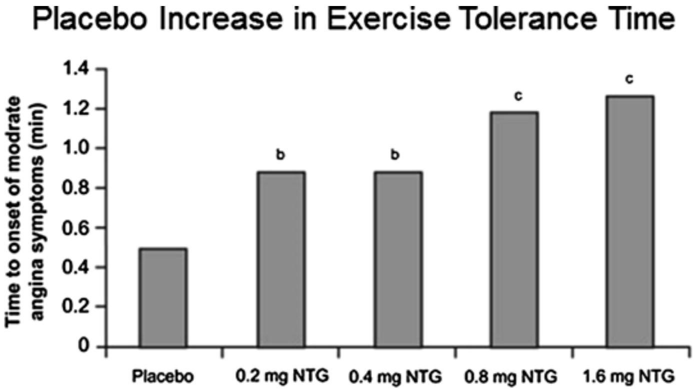


Fig. 7.5 The placebo and nocebo effect. From: Thadani and Wittig [34]

There is a problem inherent in all modern trials of antianginal therapy: because anginal patterns vary and, with modern treatments, are infrequent, a surrogate measure of antianginal effect has been adopted by the FDA and consists of treadmill walking time to the point of moderate angina. Also, just as there is a placebo effect on angina frequency, a patient's treadmill walking time frequently (50–75 %) improves with placebo therapy (Fig. 7.5). Other potential mechanisms also partially explain the improvement in exercise walking time in antianginal studies and are unrelated to a treatment effect: they are the 'learning phenomenon,' and the 'training effect.' Because of the learning phenomenon, patients frequently show an improvement in walking time between the first and second treadmill test in the absence of any treatment. The presumption is that the first test is associated with anxiety and unfamiliarity, which is reduced during the second test. Of greater importance is the training effect, with which the frequency of treadmill testing may result in a true improvement in exercise performance irrespective of treatment.

The effect of placebo on exercise tolerance in patients with angina was demonstrated in the Transdermal Nitroglycerin Cooperative Study [35], which analyzed various doses of transcutaneous-patch nitroglycerin administered for 24-h periods, in comparison with placebo patch treatment. This study was particularly important because it was the first large study to address the issue of nitrate tolerance with transcutaneous patch drug delivery in outpatient ambulatory patients. The result of the study was the demonstration of tolerance in all treated groups; the treated groups performed no better than the placebo group at the study's end. However, there was an equally striking improvement of 80 to 90s in the placebo and active treatment groups in the primary efficacy end point, walking time on a treadmill. This improvement in the placebo group could have masked any active treatment effect, but it also demonstrated the importance of a placebo control, because without this type of control, significant improvement could have been attributed by deduction to active therapy.

It was once thought that internal mammary artery ligation improved angina pectoris until studies showed a similar benefit in patients in whom a sham operation, consisting of skin incision with no ligation, was performed. Beecher [36] tried to analyze the effect of doctors' personalities on clinical outcomes of internal artery ligation, by comparing the results of the same placebo procedure performed by one of two groups, the 'enthusiasts' or the 'skeptics.' His analysis indicated that the enthusiasts achieved nearly four times more 'complete relief' for patients than did the skeptics, even though the procedure has no known specific effects [36]. Five patients undergoing the sham operation emphatically described marked improvement [37, 38]. In objective terms, a patient undergoing the sham operation had an increase in work tolerance from 4 to 10 min with no inversion of T waves on the electrocardiogram and no pain. The internal mammary artery ligation procedure was used in the United States for 2 years before it was discontinued, when the procedure was disproved by three small, well-planned, double-blind studies [39].

Carver and Samuels also addressed the issue of sham therapy in the treatment of coronary artery disease [40]. They pointed out that although the pathophysiologic features of coronary artery disease are well known, the awareness of many of the expressions of myocardial ischemia are subjective, rendering the placebo effect more important. This factor has resulted in several treatments that are based on testimonials rather than scientific evidence and that have been touted as 'breakthroughs.' Among therapies cited by these authors are chelation therapy, various vitamin therapies, and mineral supplements. It has been estimated that 500,000 patients per year in the United States are treated by these techniques. Before 1995, the data to support claims regarding the effectiveness of chelation therapy were obtained from uncontrolled open-label studies. In 1994, van Rij et al. performed a double-blind, randomized, placebo-controlled study in patients with intermittent claudication and demonstrated no difference in outcomes between chelation and placebo treatments [41] The evaluated variables included objective and subjective measures, and improvement in many of the measures was shown with both therapies. Again, without the use of a placebo control, the results could have been interpreted as improvement as a result of chelation treatment. Adding to the controversy, however, are the results from the chelation arm of the Trial to Assess Chelation Therapy, which showed that infusions of a form of chelation therapy using disodium ethylene diamine tetraacetic acid (EDTA) reduced cardiovascular events by 18 % compared to a placebo treatment [48]. Investigators stated that more research is needed before considering routine use of chelation therapy for all heart attack patients and it remains unapproved by the FDA. The EDTA-based chelation solution also contained high doses of vitamin C, B-vitamins, and other components [42]. In addition, the trial used a composite endpoint (see Chap. 3) and benefits were only seen in the soft endpoints of the composite. TACT also showed some other important deviations from adherence to the scientific principles of a well-controlled trial. The study randomized 1,708 patients, but 311 (18 %) were lost to follow-up, nearly all because of withdrawal of consent (289 patients), and importantly, these withdrawals were not equally distributed between the treatment groups. Significantly more patients (n = 174) withdrew from the placebo group compared with the chelation group (n = 115; hazard ratio, 0.66; *P* = .001). A similar imbalance in discontinuation from randomized treatment was observed—281 in the placebo group and 233 in the chelation group [43]. The substantial nonretention of study participants alone is sufficient to compromise the validity of the study results.

Placebo Effect in Heart Failure

In the past, the importance of the placebo effect in patients with congestive heart failure had not been recognized [49]. In the 1970s and early 1980s, administration of vasodilator therapy was given to patients in clinical trials without placebo control. Investigators believed that the cause of heart failure was predictable, so placebo-controlled trials were unnecessary. Another view of the unfavorable course of heart failure concluded that withholding a promising new agent was unethical. The ethical issues involved when placebo therapy is considered are addressed later in this chapter.

With the inclusion of placebo controls in clinical trials, a 25–35 % improvement of patients' symptoms was documented in the placebo arms of studies. This placebo response occurred in patients with mild to severe symptoms and did not depend on the size of the study. The assessment of left ventricular (LV) function can be determined by several methods, including noninvasive echocardiography, radionuclide ventriculography, or invasive pulmonary artery balloon-floatation catheterization. These methods measure the patient's response to therapy or the natural progression of the patient's heart failure [44]. Noninvasive measurements of LV ejection fraction vary, especially when the ventricular function is poor and the interval between tests is 3–6 months. Packer found that when a 5 % increase in ejection fraction was used to determine a beneficial response to a new drug, 20–30 % of patients showed improvement while receiving placebo therapy [50]. Overall, changes in noninvasive measures of LV function have not been shown to correlate closely with observed changes in the clinical status of patients with CHF. Most vasodilator and inotropic drugs can produce clinical benefit without a change in LV ejection fraction. Conversely, LV ejection fraction may increase significantly in patients who have heart failure and worsening clinical status [44].

When invasive catheterization is used to evaluate the efficacy of a new drug, interpretation must be done carefully because spontaneous fluctuations in hemodynamic variables occur in the absence of drug therapy. To avoid the attribution of spontaneous variability to drug therapy, postdrug effects should be assessed at fixed times and threshold values should eliminate changes produced by spontaneous variability. Another factor that can mimic a beneficial drug response, by favorably affecting hemodynamic measurements, is measurement performed immediately after catheterization of the right side of the heart or after ingestion of a meal. After intravascular instrumentation, systemic vasoconstriction occurs and resolves after 12–24 h. When pre-drug measurements are done during the post-catheterization period, any subsequent measurements will show beneficial effects because the original measurements were taken in the vasoconstricted state. Comparative data must be acquired after the post-catheterization vasoconstricted state has resolved [50].

In the past, one of the most common tests to evaluate drug efficacy for heart failure was the exercise tolerance test. An increased duration of exercise tolerance represents a benefit of therapy. However, this increased duration is also recorded during placebo therapy and possibly results from the familiarity of the patient with the test, as in the learning phenomenon described earlier in this chapter for antianginal therapy; and, the increased willingness of the physician to encourage the patient to exercise to exhaustion. Placebo response to repeated exercise tolerance testing can result in an increase in duration of 90–120 s, when only one or two baseline measurements are done. This response can be reduced to 10–30 s, when 3–10 baseline measurements are performed. Another interesting finding was that the magnitude of the placebo response was directly proportional to the number of investigators in the study! Attempts to eliminate the placebo response, including the use of gas exchange measurements during exercise tolerance testing, have failed [44].

Because all methods used to measure the efficacy of a treatment for heart failure include placebo effects, studies must include placebo controls to prove the efficacy of a new drug therapy. Statistical analysis of placebo-controlled studies must compare results between groups for statistical significance. 'Between groups' refers to comparison of the change in one group, such as one receiving a new drug therapy, with the change in another group, such one receiving as a placebo [44]. For example, Archer and Leier reported that placebo therapy for 8 weeks in 15 patients with CHF resulted in a mean improvement in exercise duration of 81 s, to 30 % above baseline [51]. This result was statistically significant compared with the 12-s improvement in the nine patients in the nonplacebo control group. There were no statistically significant differences between the placebo and non-placebo groups at baseline or at week 8 of treatment by between-group statistical analysis. Echocardiography showed no significant improvement in left ventricular function in either group, and no significant differences between the two groups at baseline or during the treatment period. To prove the existence of, and to quantitate the therapeutic power of placebo treatment in CHF, all studies were performed by the same principal investigator with identical study methods and conditions, and all patients were familiarized similarly with the treadmill testing procedure before baseline measurements. Also, the study used a well-matched, nonplacebo control group and this illustrated the spontaneous variability of CHF [45].

Placebo Effect in Hypertension

Some studies of the placebo response in patients with hypertension have shown a lowering of blood pressure [46–51], but others have not [52–56]. In a Medical Research Council study, when active treatment was compared with placebo therapy (given to patients with mild hypertension for several months) similar results were produced in the two groups—an initial decrease in blood pressure followed by stabilization [46]. Of historical note is a study by Goldring et al. published in 1956. These authors fabricated a sham therapeutic 'electron gun' designed to be as 'dramatic as possible, but without any known physiologic action other than a psychogenic one.' Initial exposure to 'the gun' lasted 1–3 min and was increased to 5 min three times daily. The investigators noticed substantially decreased blood pressure during therapy compared with pre-therapy. In six of nine hospitalized patients there was a systolic/diastolic blood pressure reduction of 39/28 mmHg.

An important factor to consider is the method used to measure blood pressure. With the use of standard sphygmomanometry, in hypertensive patients, blood pressure initially decreases upon multiple measurements. In other studies of BP, 24-h intraarterial pressure measurements and circadian curves did not show a decrease in blood pressure or heart rate during placebo therapy; however, Intraarterial blood pressure measurements at home were lower than measurements at the hospital. The circadian curves from intraarterial ambulatory blood pressure monitoring were reproducible on separate days, several weeks apart [57]. Similar to 24-h invasive intra-arterial monitoring, 24-h noninvasive automatic ambulatory blood pressure also is apparently devoid of a placebo effect. In one study, on initial application of the blood pressure device, a small reduction in ambulatory blood pressure values in the first 8 h occurred with placebo therapy. This effect, however, did not change the mean 24-h value. The home monitoring values were lower than the office measurements. Heart rate also was measured, with no variance in either setting. The office measurement of blood pressure was lower after 4 weeks of placebo therapy, but the 24-h blood pressure measurement was not [58]. This study confirmed the absence of a placebo effect in 24-h noninvasive ambulatory blood pressure monitoring, as suggested by several specific studies on large numbers of patients [59, 60]. The 24-h monitoring was measured by the noninvasive automatic Spacelabs 5300 device (Spacelabs, Redmond, Wash.) [61]. Another important factor in 24-h noninvasive monitoring is that the intervals of measurement were <60 min [62].

In a study on the influence of observer's expectation on the placebo effect in blood pressure measurements, 100 patients were observed for a 2-week single-blind period and for a 2-week double-blind period [63]. During this time, the patients' blood pressures were measured by two methods: a 30-min recording with an automatic oscillometric device and a standard sphygomomanometric measurement performed by a physician. All patients were seen in the same examining room and seen by the same physician and their blood pressure monitored by the same automatic oscillometric device. The results during the single-blind period showed a slight but statistically significant decrease in diastolic blood pressure detected by the automatic oscillometric device and no decrease measured by the physician. During the double-blind period, there was no additional decline in diastolic blood pressure measured by the oscillometric device, but the physician measured significant decreases in systolic and diastolic blood pressures. Overall, the blood pressures measured by the automatic oscillometric device, in the absence of the physician, were lower than those measured by the physician. However, there was significant correlation between the two methods. It should be mentioned that although there was a placebo effect in the measurement of blood pressure in the landmark Systolic Hypertension in the Elderly Program, it was not as significant as the reduction in blood pressure produced by active therapy in patients ≥60 years of age who had isolated systolic hypertension.

As was true with angina studies, questions have been raised about the safety of placebo control studies in hypertension. As a result, two recent publications have addressed this issue [38, 71]. Al-Khatib et al. performed a systematic review of the safety of placebo controls in short-term trials [70]. In their meta-analysis, they combined the data for death, stroke, MI, and CHF from 25 randomized trials. Each study was relatively small (n = 20–734) but the combined sample size was 6409. They found a difference between the two treatment groups and at the worst there were no more than 6/10000 difference between placebo and active therapy. Lipicky et al. reviewed all original case report forms for deaths and dropouts were reviewed from al anti-hypertensive drug trials submitted to the FDA (as an NDA) between 1973 and 2001 [64]. The population at risk was 86,137 randomized patients; 64,438 randomized to experimental drug, and 21,699 to placebo. Of the 9636 dropouts more were from the placebo group (RR 1.33 for placebo), the majority of the dropouts were, as expected, due to treatment failures, and the patients were simply returned to their original therapies with no sequelae. When serious adverse events were compared (death, irreversible harm, etc.) there were no differences between placebo and experimental drug.

Placebo Effect in Arrhythmia

Spontaneous variability in the natural history of disease or in its signs or symptoms is another reason that placebo controls are necessary. In a study of ventricular arrhythmias, Michelson and Morganroth found marked spontaneous variability of complex ventricular arrhythmias such as ventricular tachycardia and couplets [65]. These investigators observed 20 patients for 4-day periods of continuous electrographic monitoring. They recommended that when evaluating therapeutic agents, a comparison of one 24-h control period to four 24-h test periods must show a 41 % reduction in the mean hourly frequency of ventricular tachycardia and a 50 % reduction in the mean hourly frequency of couplets to demonstrate statistically significant therapeutic efficacy. They also suggested that individual patient data not be pooled to detect trends because individual variability was so great. In another study by Morganroth et al. an algorithm to differentiate spontaneous variability from proarrhythmia in patients with benign or potentially lethal ventricular arrhythmias was provided. Two or more Holter tracings were examined from each of 495 patients during placebo therapy. The algorithm defined proarrhythmia as a >3-fold increase in the frequency of ventricular premature complexes (VPCs) when the baseline frequency of ventricular premature complexes VPCs/h and a >10-fold increase when the frequency was <100 VPCs/h. The false-positive rate was 1 % when this algorithm was used.

The Cardiac Arrhythmia Suppression Trial (CAST) evaluated the effect of antiarrhythmic therapy in patients with asymptomatic or mildly symptomatic ventricular arrhythmia [66]. Response to drug therapy was determined by a ≥80 % reduction in ventricular premature depolarizations or a ≥90 % reduction in runs of unsustained ventricular tachycardia as measured by 24-h Holter monitoring 4–10 days after initiation of pharmacologic treatment, a response previously considered to be an important surrogate measure of antiarrhythmic drug efficacy. One thousand four hundred fifty-five patients were assigned to drug regimens, and ambulatory electrocardiographic (Holter) recording screened for arrhythmias. The CAST Data and Safety Monitoring Board recommended that encainide and flecainide therapy be discontinued because of the increased number of deaths from arrhythmia, cardiac arrest, or any cause compared with placebo treatment. The CAST investigators conclusion emphasized the need for more placebo-controlled clinical trials of antiarrhythmic drugs with a mortality end point.

Relation of Treatment Adherence to Survival in Patients with or Without History of Myocardial Infarction

An important consideration in determining study results is adherence to therapy and the presumption that any differences in adherence rates would be equal in the active versus the placebo treatment groups. The Coronary Drug Project Research Group [67] planned to evaluate the efficacy and safety of several lipid-influencing drugs in the long-term treatment of coronary heart disease. This randomized, double-blind, placebo-controlled, multicenter clinical trial found no significant difference in the 5-year mortality of 1,103 men treated with the fibric acid derivative clofibrate compared with 2,789 men given placebo. However, subjects showing good adherence (patients taking ≥80 % of the protocol drug) had lower mortality than did subjects with low adherence in both the clofibrate group and the placebo group [67].

A similar association between adherence and mortality was found in patients after myocardial infarction in the Beta-Blocker Heart Attack Trial data [72]. This phenomenon was extended to women after myocardial infarction. On analysis of the trial data for 505 women randomly assigned to β-blocker therapy or placebo therapy, there was a 2–2.5-fold increase in mortality within the first 2 years in patients taking <75 % of their prescribed medication. Adherence among men and women was similar, at about 90 %. However, the cause of the increased survival resulting from good adherence is not known. There is speculation that good adherence reflects a favorable psychological profile—a personal ability to make lifestyle adjustments that limit disease progression. Alternatively, adherence may be associated with other advantageous health practices or social circumstances not measured. Another possible explanation is that improved health status may facilitate good adherence [68].

The Lipid Research Clinics Coronary Primary Prevention Trial [69] did not find a correlation between compliance and mortality. These investigators randomly assigned 3806 asymptomatic hypercholesterolemic men to receive cholestyramine or placebo. The main effects of the drug compared with placebo on cholesterol level and death or nonfatal myocardial infarction were analyzed over a 7-year period. In the group receiving active drug, a relation between compliance and outcome existed, mediated by a lowering of cholesterol level. However, no effect of compliance on cholesterol level or outcome was observed in the placebo group [69, 70].

The Physicians' Health Study included a randomized fashion 22,000 United States male physicians 40–84 years old who were free of myocardial infarction and cerebral vascular disease [71]. This study analyzed the benefit of differing frequencies of aspirin consumption on the prevention of myocardial infarction. In addition, the study identified factors associated with adherence and analyzed the relation of adherence with cardiovascular outcomes in the placebo group. Analysis showed an average compliance of 80 % in the aspirin and placebo groups during the 60 months of follow-up [71]. Adherence during that trial was associated with several baseline characteristics in both the aspirin and placebo groups as follows. Trial participants with poor adherence (<50 % compliance with pill consumption), relative to those with good adherence, were more likely to be younger than 50 years at randomization, to smoke cigarettes, to be overweight, not to exercise regularly, to have a parental history of myocardial infarction, and to have angina. These associations were statistically significant. In a multivariate logistic regression model, cigarette smoking, excess weight, and angina remained significant predictors of poor compliance. The strongest predictor of adherence during the trial was adherence during the run-in period. Baseline characteristics with little relation to adherence included regular alcohol consumption and a history of diabetes and hypertension [71]. Using intention-to-treat analysis, the aspirin group had a 41 % lower risk of myocardial infarction compared with the placebo group. On subgroup analysis, participants reporting excellent (≥95 %) adherence in the aspirin group had a significant, 51 % reduction in the risk of first myocardial infarction relative to those with similar adherence in the placebo group. Lower adherence in the aspirin group was not associated with a statistically significant reduction in first myocardial infarction compared with excellent adherence in the placebo group. Excellent adherence in the aspirin group was associated with a 41 % lower relative risk of myocardial infarction compared with low adherence in the aspirin group. Excellent adherence in the placebo group was not associated with a reduction in relative risk. The rate of stroke was different from that of myocardial infarction. On intention-to-treat analysis, the aspirin group had a nonsignificant, 22 % increased rate of stroke compared with the placebo group. Participants with excellent adherence in the placebo group had a lower rate of strokes than participants in the aspirin or placebo groups with low (<50 %) adherence. Excellent adherence in the placebo group was associated with a 29 % lower risk of stroke compared with excellent adherence in the aspirin group.

Also analyzed in the above study, was the overall relation of adherence to aspirin therapy with cardiovascular risk when considered as a combined end point of all important cardiovascular events, including first fatal or nonfatal myocardial infarction or stroke or death resulting from cardiovascular disease with no previous myocardial infarction or stroke. On intention-to-treat analysis, there was an 18 % decrease in the risk of all important cardiovascular events in the aspirin group compared with the

Table 7.4 Placebo adherence and mortality

Outcome	HR for adherence
Total mortality	.52
CVD mortality	.66
Non CVD mortality	.40
CHD mortality	.54
Incident cancer	.42

Padula et al. [72]

placebo group. Participants with excellent adherence in the aspirin group had a 26 % reduction in risk of a first major cardiovascular event compared with those with excellent adherence in the placebo group. However, participants in the aspirin group with low compliance had a 31 % increased risk of a first cardiovascular event compared with those in the placebo group with excellent adherence. Within the placebo group, there was no association between level of adherence and risk of a first cardiovascular event. In the analysis of death resulting from any cause in persons with a previous myocardial infarction or stroke, low adherence in both the aspirin group and the placebo group was associated with a fourfold increase in the risk of death. When the 91 deaths due to cardiovascular causes were studied, similar elevations in risk were found in both the placebo and aspirin groups with poor adherence compared with those in the placebo group with excellent adherence.

The Physicians' Health Study [71] found results similar to those of the Coronary Drug Project when all cause mortality and cardiovascular mortality were considered [67]. These relations remained strong when adjusted for potential confounding variables at baseline. The strong trend for higher death rates among participants with low adherence in both the aspirin and the placebo groups may be due to the tendency for subjects to decrease or discontinue study participation as their health declines to serious illness. Low adherence in the placebo group was not associated with an increased risk of acute events such as myocardial infarction. Thus placebo effects seem to vary depending on the outcome considered.

Most recently has been an analysis of the Hormone Estrogen Replacement Study, a secondary prevention study of CHD in postmenopausal women (Table 7.4) [30]. Investigators also evaluated the association of placebo adherence and total mortality and found that the more adherent participants had significantly lower mortality than non-adherers HR 0.52 (0.29; 0.93) [72]. They speculated about the possibilities for that observation and suggested that adherence could be a marker for healthier lifestyles and/or that as a fatal illness prodrome, adherence may decrease (an effect-cause artifact).

Miscellaneous

Flaten conducted an experiment in which he told participants that they were receiving either a relaxant, stimulant, or an inactive agent, but in fact gave all of them the inactive agent. Patients who were told they were getting the relaxant showed reduced stress levels, while those who thought they were receiving the stimulant showed increased arousal levels. In another study, asthmatics that were told they were getting either a bronchodilator or bronchconstrictor and who actually received that particular therapy, had more effective responses when the information received actually matched the drug effect.

Linde et al. evaluated the placebo effect of pacemaker implantation in 81 patients with obstructive hypertrophic cardiomyopathy [78]. The study design was a 3-month multicenter, double-blind, cross-over study. In the first study period 40 patients were assigned to inactive pacing, and were compared to 41 patients with active pacing. During inactive pacing, there was an improvement in chest pain, dyspnea, palpitations, and in the left ventricular outflow gradient. The change in the active pacing group for most parameters was greater.

Clinical Trials and the Ethics of Using Placebo Controls

Since the 1962 amendments to the Food, Drug, and Cosmetic Act, the FDA has had to rely on the results of 'adequate and well-controlled' clinical trials to determine the efficacy of new pharmacologic therapies. Regulations govern pharmacologic testing and recognize several types of controls that may be used in clinical trials to assess the efficacy of new pharmacologic therapies. The controls include (1) placebo concurrent control, (1) dose-comparison concurrent control, (2) no-treatment concurrent control, (4) active-treatment concurrent control, and (5) historical control (Table 7.5). Regulations, however, do not specify the circumstances for the use of these controls because there are various study designs that may be adequate in a given set of circumstances [18].

There is ongoing debate concerning the ethics of using placebo controls in clinical trials of cardiac medications. The issue revolves around the administration of placebo in lieu of a proven therapy. Two articles, by Rothman and Michels [73] and Clark and Leaverton [74], illustrate the debate. Rothman and Michels [73] state that patients in clinical trials often receive placebo therapy instead of proven therapy for the patient's medical condition and assert that this practice is in direct violation of the Nuremberg Code and the World Medical Association's adaptation of this Code in the Declaration of Helsinki. The Nuremberg Code, a 10-point ethical code for experimentation in human beings, was formulated in response to the human experimentation atrocities that were recorded during the post-World War II trial of Nazi physicians in Nuremberg, Germany. According to Rothman and Michels [73],

Table 7.5 Types of treatment controls

Placebo concurrent control
Dose-comparison concurrent control
No-treatment concurrent control
Active-treatment concurrent control
Historical control

violation occurs because the use of placebos as controls denies the patient and best proven therapeutic treatment. It occurs despite the establishment of regulatory agencies and institutional review boards, although these authors seem to ignore that informed consent is part of current practice, as certainly was not the case with the Nazi atrocities. However, a survey of federally funded grants found that despite the process of informed consent almost 25 % of medical research subjects were unaware that they were even part of a research project or that they were receiving investigational therapies. It should be noted, however, that this survey spanned 20 years, and did not include analysis for the more recent time period, when, most would agree, there has been more emphasis on informed consent.

One reason why placebo-controlled trials are approved by institutional review boards is that this type of trial is part of the FDA's general recommendation for demonstrating therapeutic efficacy before an investigational drug can be approved. That is, according to the FDA, when an investigational drug is found to be more beneficial by achieving statistical significance over placebo therapy, then therapeutic efficacy is proven [75]. As more drugs are found to be more effective than placebos in treating diseases, the inclusion of a placebo group is often questioned. However, this question ignores that in many cases drug efficacy in the past had been established by surrogate measures; and, as new and better measures of efficacy become available, additional study becomes warranted. Regarding surrogate measures and their potential to mislead, the study of the suppression of ventricular arrhythmia by antiarrhythmic therapy was later proven to be unrelated to survival; in fact, results with active therapy were worse than with placebo. Likewise, in studies of inotropic therapy for heart failure, exercise performance rather than survival was used as the measure of efficacy, when in fact a presumed efficacious therapy performed worse than placebo when survival was assessed. In the use of immediate short-acting dihydropyridine calcium antagonist therapy for the relief of symptoms of chronic stable angina pectoris, again a subject might have fared better had he or she been randomly assigned to placebo therapy.

Also important to the concept that established beneficial therapy should not necessarily prohibit the use of placebo in the evaluation of new therapies is that the natural history of a disease may change, and the effectiveness of so-called established therapies (e.g., antibiotic agents for treatment of infections) may diminish. When deciding on the use of an investigational drug in a clinical trial, the prevailing standard is that there should be enough confidence to risk exposure to a new drug, but enough doubt about the drug to risk exposure to placebo. Thus, in this situation, the use of a placebo control becomes warranted, particularly as long as other live-saving therapy is not discontinued.

The use of placebo-controlled trials may be advocated on the basis of a scientific argument. When pharmacologic therapy was shown to be effective in previous placebo-controlled trials, conclusions made from current trials without placebo controls may be misleading because the previous placebo-controlled trial then becomes a historical control. Historical controls are the least reliable for demonstration of efficacy [18]. In active-controlled clinical trials without a placebo arm, there is an assumption that the active control treatment is as effective under the new experimental conditions as it was in the previous placebo-controlled clinical trial. This assumption can result in misleading conclusions when results with an experimental therapy are found to be equivalent to those with the active, proven therapy. This conclusion of equivalence can be magnified by conservative statistical methods, such as the use of the 'intent-to-treat' approach, an analysis of all randomized patients regardless of protocol deviations, and an attempt to minimize the potential for introduction of bias into the study. Concurrent placebo controls account for factors other than drug-effect differences between study groups. When instead of a placebo-control group an untreated control group is used, then blinding is lost and treatment-related bias may occur [18, 74].

Clark and Leaverton [74] and Rothman and Michels [73] agree that the use of placebo controls is ethical when there is no existing treatment to favorably affect morbidity and mortality. Furthermore, there are chronic diseases for which treatment exists that not favorably alter morbidity and mortality. For example, no clinical trial has found the treatment of angina to increase a patient's survival. In contrast, treatment after a myocardial infarction with β-blocking agents has been convincingly proven to increase a patient's survival [74]. However, Clark and Leaverton [74] disagree with Rothman and Michels [73] in that they assert that for chronic disease, a placebo-controlled clinical trial of short duration is ethical because there is usually no alteration in long-term outcome for the patient. The short duration of the trial represents a small segment of the lifetime management of a chronic disease. For instance, the treatment of chronic symptomatic CHF and a low ejection fraction (<40 %) with enalapril was shown to decrease mortality by 16 %. This decrease in mortality was most marked in the first 24 months of follow-up, with an average follow-up period of 40 months. Therefore, only long-term compliance with pharmacologic therapy resulted in some decreased mortality. Another example of a chronic medical condition that requires long-term treatment and in which short-term placebo is probably not harmful is hypertension [76]. In some studies men and women with a history of myocardial infarction and with a ≥80 % compliance with treatment, including placebo therapy, had an increased survival. This increased survival was also described in patients in a 5-year study of the effects of lipid-influencing drugs on coronary heart disease. [67, 68, 77].

A different argument for the ethical basis of using placebo controls relies on the informed consent process. Before a patient's participation in a clinical trial, the patient is asked to participate in the trial. The informed consent process includes a description of the use of placebos along with other aspects of the trial. In this written agreement, the patient is responsible for notifying the physician of any medical problems and is informed of his or her right to withdraw from the study at any time, as described in the Nuremberg Code and the Declaration of Helsinki. During this disclosure, patients are presented with the risks and benefits of the study. On the basis of this information, a patient voluntarily decides to participate, knowing that he or she may receive a placebo or investigational medication.

All parties involved in research should be responsible for their research and accountable for its ethics. Clinical trials failing to comply with the Nuremberg Code and the Declaration of Helsinki should not be conducted and should not be accepted for publication. Yet, there is disagreement in determining which research methods are in compliance with the Nuremberg Code and Declaration of Helsinki. Scientific needs should not take precedence over ethical needs. Clinical trials need to be carefully designed to produce a high quality of trial performance. In addition, in experimentation involving human subjects, the Nuremberg Code and Declaration of Helsinki must be used as universal standards. The Declaration of Helsinki addresses the selection of appropriate controls by stating 'the benefits, risks, burdens, and effectiveness of a new method should be tested against the best current prophylactic, diagnostic, and therapeutic methods. This does not exclude the use of placebo, or of no treatment, in studies where no proven prophylactic, diagnostic, or therapeutic method exists.' Others have added that if the patient or subject is not likely to be harmed through exposure to placebo, and they can give voluntary informed consent, it is permissible to use placebo controls in some trials despite the existence of a know effective therapy.

Conclusions

Until the mechanism of the placebo action is understood and can be controlled, a clinical trial that does not include a placebo group provides data that should be interpreted with caution. The absence of a placebo group makes it difficult to assess the true efficacy of a therapy. It is easy to attribute clinical improvement to a drug therapy when there is no control group. As was found with heart failure, almost all chronic diseases have variable courses. In addition, because each clinical trial has a different setting and different study design within the context of the physician-patient relationship, a placebo group helps the investigator differentiate true drug effects from placebo effects.

More important than the inclusion of a placebo group is a careful study design that includes frequent review, by a data and safety monitoring board, of each patient's medical condition. This monitoring is crucial to protect the study participants. To protect the participants, trials must include provisions that require a patient to be removed from a trial when the patient or doctor believes that removal is in the patient's best interest. The patient can then be treated with currently approved therapies.

Patients receiving placebo may report subjective clinical improvements, and demonstrate objective clinical improvement, for instance on exercise tolerance testing or Holter monitoring of ischemic events. Findings such as these dispel the implication that placebo therapy is the same as no therapy and may occur because many factors are involved in the physician-patient relationship such as the psychological state of the patient; the patient's expectations and conviction in the efficacy of the method of treatment' and the physician's biases, attitudes, expectations, and methods of communication [2]. An explanation of improvement in patients participating in trials is the close attention received by patients from the investigators. Baseline laboratory values are checked to ensure the safety of the patient and compliance with the study protocol. This beneficial response by the patient is called a positive placebo effect when found in control groups of patients receiving placebo therapy [30, 33, 36, 37, 39, 44, 63, 78].

Conversely, the condition of patients receiving placebos has also in some cases worsened. Every drug has side effects. These side effects are also found with placebo therapy and can be so great that they preclude the patient's continuation with the therapy. This phenomenon is always reported by patients in clinical trials receiving placebo [14, 32, 44, 63, 79, 80]. Finally, placebos can act synergistically and antagonistically with other specific and nonspecific therapies. Therefore much is still to be discovered about the placebo effect.

The arguments in support of the use of placebo controls (placebo "orthodoxy") are numerous. The word "orthodoxy" is from the Greek ortho ('right', 'correct') and doxa ('thought', 'teaching', 'glorification'). Orthodoxy is typically used to refer to the correct theological or doctrinal observance of religion, as determined by some overseeing body. The term did not conventionally exist with any degree of formality (in the sense in which it is now used) prior to the advent of Christianity in the Greek-speaking world, though the word does occasionally show up in ancient literature in other, somewhat similar contexts. Orthodoxy is opposed to heterodoxy ('other teaching'), heresy and schism. People who deviate from orthodoxy by professing a doctrine considered to be false are most often called heretics. Some of the supporting arguments are that there are methodologic limitations of trials using active controls such as:

– Variable responses to drugs in some populations
– Unpredictable and small effects
– Spontaneous improvements

In addition, some believe that no drug should be approved unless it is clearly superior to placebo or no treatment, so that placebo is ethical if there is "no permanent adverse consequence" form its use; or, if there is "risk of only temporary discomfort", or if there "is no harm" consequent to its use. It should be noted that these latter two arguments are not equivalent; that is, patients may be harmed by temporary but reversible conditions, and that these criteria may in fact permit intolerable suffering. For example, in the 1990s several placebo-controlled trials of ondansetron for chemotherapy induced vomiting were performed when there were existent effective therapies (i.e. no permanent disability, but more than mere discomfort). Another example might be the use of placebo-controlled trials of antidepressants, in which there might occur instances of depression-induced suicide.

Others argue for the use of active-controls (Active-control "Orthodoxy") in lieu of placebo controls. They argue that whenever an effective intervention for a condition exists, it must be used as the control group; that is, the clinically relevant question is not whether a new drug is better than nothing, but whether it is better than standard treatment. The supporters of the use of active controls point to the most recent "Declaration of Helsinki" which states; "the benefits, risks, burdens, and effectiveness of a new method should be tested against those of the most current prophylactic, diagnostic, or therapeutic methods. This does not exclude the use of placebo, or no treatment, in studies where no proven prophylactic, diagnostic or therapeutic method exists."

The problem with "Active-Control Orthodoxy" is that scientific validity constitutes a fundamental ethical protection, and that scientifically invalid research cannot be ethical no matter how safe the study participants are. Thus, the almost absolute prohibition of placebo in every case in which an effective treatment exists is too broad, and that patients exposed to placebo may be better off than the group exposed to a new intervention. These authors agree with Emmanual and Miller in support of a "middle ground" as discussed above.

Summary

The effect of placebo on the clinical course of systemic hypertension, angina pectoris, silent myocardial ischemia, CHF, and ventricular tachyarrhythmia's has been well described. In the prevention of myocardial infarction, there appears to be a direct relation between compliance with placebo treatment and favorable clinical outcomes. The safety of short-term placebo-controlled trials has now been well documented in studies of drug treatment of angina pectoris. Although the ethical basis of performing placebo-controlled trials continues to be challenged in the evaluation of drugs for treating cardiovascular disease, as long as a life-saving treatment is not being denied it remains prudent to perform placebo-controlled studies for obtaining scientific information.

References

1. Turner JA, Deyo RA, Loeser JD, Von Korff M, Fordyce WE. The importance of placebo effects in pain treatment and research. JAMA. 1994;271:1609–14.
2. Benson H, Epstein MD. The placebo effect. A neglected asset in the care of patients. JAMA. 1975;232:1225–7.
3. Stedman's Medical Dictionary. Melmon and Morrelli's clinical pharmacology: basic principles in therapeutics. In: Melmon K, Morrelli H, Hoffman B, Mierenberg D, editors. 3rd ed. New York: McGraw-Hill; 1992. p. 896.
4. Packer M, Medina N, Yushak M. Hemodynamic changes mimicking a vasodilator drug response in the absence of drug therapy after right heart catheterization in patients with chronic heart failure. Circulation. 1985;71:761–6.
5. Chalmers TC. Prophylactic treatment of Wilson's disease. N Engl J Med. 1968;278:910–1.
6. Garrison FH. History of medicine. 4th ed. Philadelphia: Saunders; 1929.
7. Stedman's Medical Dictionary. Baltimore: Williams & Wilkins; 1990. Stedman's Medical Dictionary.
8. White L, Tursky B, Schwartz G. Placebo: theory, research, and mechanisms. New York: Guilford Press; 1985. 474 p.
9. Shapiro AK. Factors contributing to the placebo effect: their implications for psychotherapy. Am J Psychother. 1961;18:73–88.
10. Byerly H. Explaining and exploiting placebo effects. Perspect Biol Med. 1976;19:423–36.
11. Lind JA. A treatise of the scurvy. Edinburgh;1753.
12. Hill AB. The clinical trial. Br Med Bull. 1951;7:278–82.
13. Beecher HK. The powerful placebo. JAMA. 1955;159:1602–6.
14. Lasagna L, Mosteller F, Von Felsinger JM, Beecher HK. A study of the placebo response. Am J Med. 1954;16:770–9.

15. Wolf S, Pinsky RH. Effects of placebo administration and occurrence of toxic reactions. JAMA. 1954;155:339–41.
16. Davis JM. Don't let placebos fool you. Postgrad Med. 1990;88:21–4.
17. Nies A, Spielberg S. Principles of therapeutics. In: Hardman JG, Limbird LE, editors. Goodman and Gilman's the pharmacological basis of therapeutics. 9th ed. New York: McGraw Hill; 1996.
18. Makuch RW, Johnson MF. Dilemmas in the use of active control groups in clinical research. IRB. 1989;11:1–5.
19. Ederer F. Serum cholesterol changes: effects of diet and regression toward the mean. J Chronic Dis. 1972;25:277–89.
20. The National Diet-Heart Study Final Report. Circulation. 1968;37:I1–428.
21. Yudkin PL, Stratton IM. How to deal with regression to the mean in intervention studies. Lancet. 1996;347:241–3.
22. Asmar R, Safar M, Queneau P. Evaluation of the placebo effect and reproducibility of blood pressure measurement in hypertension. Am J Hypertens. 2001;14:546–52.
23. Oh VM. Magic or medicine? Clinical pharmacological basis of placebo medication. Ann Acad Med Singapore. 1991;20:31–7.
24. Kelly JP. Anatomical organization of the nervous system. In: Kandel ER, Schwartz JH, Jessel TM, editors. Principles of neural science. 3rd ed. New York: Elsevier; 1991. p. 276–92.
25. Voudouris NJ, Peck CL, Coleman G. The role of conditioning and verbal expectancy in the placebo response. Pain. 1990;43:121–8.
26. Levine JD, Gordon NC, Bornstein JC, Fields HL. Role of pain in placebo analgesia. Proc Natl Acad Sci U S A. 1979;76:3528–31.
27. Hersh EV, Ochs H, Quinn P, MacAfee K, Cooper SA, Barasch A. Narcotic receptor blockade and its effect on the analgesic response to placebo and ibuprofen after oral surgery. Oral Surg Oral Med Oral Pathol. 1993;75:539–46.
28. Kojo I. The mechanism of the psychophysiological effects of placebo. Med Hypotheses. 1988;27:261–4.
29. Egbert LD, Battit GE, Welch CE, Bartlett MK. Reduction of postoperative pain by encouragement and instruction of patients. A study of doctor-patient rapport. N Engl J Med. 1964;270:825–7.
30. Amsterdam EA, Wolfson S, Gorlin R. New aspects of the placebo response in angina pectoris. Am J Cardiol. 1969;24:305–6.
31. Glasser SP, Clark PI, Lipicky RJ, Hubbard JM, Yusuf S. Exposing patients with chronic, stable, exertional angina to placebo periods in drug trials. JAMA. 1991;265:1550–4.
32. Boissel JP, Philippon AM, Gauthier E, Schbath J, Destors JM. Time course of long-term placebo therapy effects in angina pectoris. Eur Heart J. 1986;7:1030–6.
33. McGraw BF, Hemberger JA, Smith AL, Schroeder JS. Variability of exercise performance during long-term placebo treatment. Clin Pharmacol Ther. 1981;30:321–7.
34. Thadani U and Wittig T. A Randomized, Double-Blind, Placebo-Controlled, Crossover, Dose-Ranging Multicenter Study to Determine the Effect of Sublingual Nitroglycerin Spray on Exercise Capacity in Patients with Chronic Stable Angina. Clin Med Insights Cardiol. 2012;6: 87–95.
35. Acute and chronic antianginal efficacy of continuous twenty-four-hour application of transdermal nitroglycerin. Steering Committee, Transdermal Nitroglycerin Cooperative Study. Am J Cardiol. 1991;68:1263–73.
36. Beecher HK. Surgery as placebo. A quantitative study of bias. JAMA. 1961;176:1102–7.
37. Diamond EG, Kittle CF, Crockett JE. Evaluation of internal mammary artery ligation and sham procedures in angina pectoris. Circulation. 1958;18:712–3.
38. Diamond EG, Kittle CF, Crockett JE. Comparison of internal mammary artery ligation and sham operation for angina pectoris. Am J Cardiol. 1960;5:484–6.
39. Cobb LA. Evaluation of internal mammary artery ligation by double-blind technic. N Engl J Med. 1989;260:1115–8.

40. Carver JR, Samuels F. Sham therapy in coronary artery disease and atherosclerosis. Pract Cardiol. 1988;14:81–6.
41. van Rij AM, Solomon C, Packer SG, Hopkins WG. Chelation therapy for intermittent claudication. A double-blind, randomized, controlled trial. Circulation. 1994;90:1194–9.
42. Lamas GA, Goertz C, Boineau R, Mark DB, Rozema T, Nahin RL, et al. Effect of disodium EDTA chelation regimen on cardiovascular events in patients with previous myocardial infarction: the TACT randomized trial. JAMA. 2013;309:1241–50. doi:10.1001/jama.2013.2107.
43. Nissan S. Concerns about reliability in the Trial to Assess Chelation Therapy. JAMA. 2013;309:1293–4. doi:10.1001/jama.2013.2778.
44. Packer M. The placebo effect in heart failure. Am Heart J. 1990;120:1579–82.
45. Archer TP, Leier CV. Placebo treatment in congestive heart failure. Cardiology. 1992;81:125–33.
46. Randomised controlled trial of treatment for mild hypertension: design and pilot trial. Report of Medical Research Council Working Party on Mild to Moderate Hypertension. BMJ. 1977;1:1437–40.
47. Gould BA, Mann S, Davies AB, Altman DG, Raftery EB. Does placebo lower blood-pressure? Lancet. 1981;2:1377–81.
48. Martin MA, Phillips CA, Smith AJ. Acebutolol in hypertension—double-blind trial against placebo. Br J Clin Pharmacol. 1978;6:351–6. PMC1429467.
49. Moutsos SE, Sapira JD, Scheib ET, Shapiro AP. An analysis of the placebo effect in hospitalized hypertensive patients. Clin Pharmacol Ther. 1967;8:676–83.
50. Myers MG, Lewis GR, Steiner J, Dollery CT. Atenolol in essential hypertension. Clin Pharmacol Ther. 1976;19:502–7.
51. Pugsley DJ, Nassim M, Armstrong BK, Beilin L. A controlled trial of labetalol (Trandate), propranolol and placebo in the management of mild to moderate hypertension. Br J Clin Pharmacol. 1979;7:63–8. PMC1429617.
52. A DOUBLE blind control study of antihypertensive agents. I. Comparative effectiveness of reserpine, reserpine and hydralazine, and three ganglionic blocking agents, chlorisondamine, mecamyamine, and pentolinium tartrate. Arch Intern Med. 1960;106:81–96.
53. Effects of treatment on morbidity in hypertension: results in Patients with diastolic blood pressures averaging 115 through 119 mmHg by Veterans Administration cooperative study group on antihypertensive agents. JAMA. 1967;202:116–22.
54. Effects morbidity of treatment on in hypertension II. Results in patients with diastolic blood pressure averaging 90 through 114 mm Hg. JAMA. 1970;213(7):1143–52. doi:10.1001/jama.1970.03170330025003.
55. Hansson L, Aberg H, Karlberg BE, Westerlund A. Controlled study of atenolol in treatment of hypertension. BMJ. 1975;2:367–70.
56. Wilkinson PR, Raftery EB. A comparative trial of clonidine, propranolol and placebo in the treatment of moderate hypertension. Br J Clin Pharmacol. 1977;4:289–94.
57. Raftery EB, Gould BA. The effect of placebo on indirect and direct blood pressure measurements. J Hypertens Suppl. 1990;8:S93–100.
58. Mutti E, Trazzi S, Omboni S, Parati G, Mancia G. Effect of placebo on 24-h non-invasive ambulatory blood pressure. J Hypertens. 1991;9:361–4.
59. Dupont AG, Van der Niepen P, Six RO. Placebo does not lower ambulatory blood pressure. Br J Clin Pharmacol. 1987;24:106–9.
60. O'Brien E, Cox JP, O'Malley K. Ambulatory blood pressure measurement in the evaluation of blood pressure lowering drugs. J Hypertens. 1989;7:243–7.
61. Casadei R, Parati G, Pomidossi G, Groppelli A, Trazzi S, Di Rienzo M, et al. 24-hour blood pressure monitoring: evaluation of Spacelabs 5300 monitor by comparison with intra-arterial blood pressure recording in ambulant subjects. J Hypertens. 1988;6:797–803.
62. Portaluppi F, Strozzi C, degli Uberti E, Rambaldi R, Trasforini G, Margutti A, et al. Does placebo lower blood pressure in hypertensive patients? A noninvasive chronobiological study. Jpn Heart J. 1988;29:189–97.

63. Sassano P, Chatellier G, Corvol P, Menard J. Influence of observer's expectation on the placebo effect in blood pressure trials. Curr Ther Res. 1987;41:304–12.
64. Lipicky R, DeFelice A, Gordon M, Hung J, Karkowsky A, Lawrence J, et al. Placebo in Hypertension Adverse Reaction Meta-Analysis (PHARM). Circulation. 2003;17:IV–452.
65. Michelson EL, Morganroth J. Spontaneous variability of complex ventricular arrhythmias detected by long-term electrocardiographic recording. Circulation. 1980;61:690–5.
66. Pratt CM, Moye LA. The cardiac arrhythmia suppression trial. Casting suppression in a different light. Circulation. 1995;91:245–7.
67. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project research group. N Engl J Med. 1980;303:1038–41.
68. Gallagher EJ, Viscoli CM, Horwitz RI. The relationship of treatment adherence to the risk of death after myocardial infarction in women. JAMA. 1993;270:742–4.
69. Coronary Drug Project Research Group. The Lipid Research Clinics Coronary Primary Prevention Trial results. II. The relationship of reduction in incidence of coronary heart disease to cholesterol lowering. JAMA. 1984; 251:365–74.
70. Sackett DL, Haynes RB, Gibson E, Johnson A. The problem of compliance with antihypertensive therapy. Pract Cardiol. 1976;2:35–9.
71. Glynn RJ, Buring JE, Manson JE, LaMotte F, Hennekens CH. Adherence to aspirin in the prevention of myocardial infarction. The Physicians' Health Study. Arch Intern Med. 1994;154:2649–57.
72. Padula AM, Pressman AR, Vittinghoff E, Grady D, Neuhaus J, Ackerson L, et al. Placebo adherence and mortality in the Heart and Estrogen/Progestin Replacement Study. Am J Med. 2012;125:804–10. PMC3423204.
73. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. N Engl J Med. 1994;331:394–8.
74. Clark PI, Leaverton PE. Scientific and ethical issues in the use of placebo controls in clinical trials. Ann Rev Public Health. 1994;15:19–38.
75. Schechter C. The use of placebo controls. N Engl J Med. 1995;332; author reply 2.
76. Alderman MH. Blood pressure management: individualized treatment based on absolute risk and the potential for benefit. Ann Intern Med. 1993;119:329–35.
77. Horwitz RI, Viscoli CM, Berkman L, Donaldson RM, Horwitz SM, Murray CJ, et al. Treatment adherence and risk of death after a myocardial infarction. Lancet. 1990;336:542–5.
78. Morganroth J, Borland M, Chao G. Application of a frequency definition of ventricular proarrhythmia. Am J Cardiol. 1987;59:97–9.
79. Drici MD, Raybaud F, De Lunardo C, Iacono P, Gustovic P. Influence of the behaviour pattern on the nocebo response of healthy volunteers. Br J Clin Pharmacol. 1995;39:204–6.
80. Roberts AH. The powerful placebo revisited: magnitude of nonspecific effects. Mind/Body Med. 1995;1:35–43.
81. Glasser SP, Willard J, Defelice A, Lawrence J, Hung J, Obot E, et al. Is randomization to placebo safe? Risk in placebo-controlled angina trials: angina risk meta-analysis. Cardiology. 2011; 120(2):174–81. Epub 2012/01/21.

Chapter 8
Recruitment and Retention in Clinical Research

Stephen P. Glasser

Abstract Nothing is more important to a clinical research study than recruiting and then retaining subjects in a study. In addition, losses to follow-up can destroy a study. This chapter will address issues such as why people participate in clinical research, what strategies can be employed to recruit and then retain subjects in a study, issues involved with minority recruitment, and HIPAA; and, will include some real examples chosen to highlight the retention of potential drop-outs.

Keywords Recruitment process • Target population • Recruitment failure • Minority recruitment • Opt in-opt out • Participation • Types of recruitment

Nothing is more important to a clinical research study than recruiting and then retaining subjects in a study. However, many studies fail to recruit their planned number of participants. For example, between 1997 and 2009, US federally funded RCTs of coronary artery disease populations required a significant number of non-US patients for the study to achieve enrollment goals [1]. In the United Kingdom, more than 2 out of 3 trials failed to recruit the target population within the studies pre-stated time frame [2]. In an accompanying editorial to the aforementioned study, it was suggested that the success in recruitment to randomized trials comes down to "keep it simple and close to home…". Of course, studies that recruit too few patients might miss clinically important effects, so the scale of the problem has been assessed; and, in one study that consisted of a multi-center cohort trial, only 37 % of the trials met their planned recruitment goals [3]. Easterbrook et al. also studied the issue of recruitment in 487 research protocols submitted to the Central Oxford Research Ethics Committee, and found that 10 never started, and 16 reported abandonment of the study, because of recruitment difficulties [4]. When it comes to

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_8,

regulatory approval of a new therapeutic entity, controlling variability is essential since the patient population being studied not only determines the potential market size for a drug, but also impacts the drug's efficacy conclusion. The patient's status can impact variability, and that is why proper patient recruitment is so critical. At the other end of the spectrum is that too often patient recruitment is treated as a numbers game, and as such it may result in enrollment of patients whose Inclusion/Exclusion criteria are "stretched" or who may not be reliable enough to complete the trial. To address this latter issue, Blinded Independent Central Reviews (BICR) is becoming more common, and is generally handled by the use of a Contract Research Organization (CRO). The importance of recruiting patients/subjects into a research study who will maintain good compliance and who will complete the study cannot be overemphasized, because patient discontinuations in a research study decreases the studies power, validity and potential treatment effect.

Not only is patient recruitment a critical issue, but patient retention is equally critical. Particularly, losses to follow-up can destroy a study (see Chap. 3). Recruitment and retention has become even more important in today's environment of scandals, IRB constraints, HIPPA, the ethics of reimbursing study participants, and skyrocketing costs. For example, one researcher demonstrated how not to do research as outlined in a USA Today article in 2000 [5]. According to that newspaper article, the researcher put untoward recruitment pressure on the staff, ignored other co-morbid diseases in the recruited subjects, performed multiple simultaneous studies in the same subjects, fabricated and destroyed records, and ultimately blamed the study coordinators for all the errors found during an audit.

Recruitment Process

Human clinical trials are critical to improving medical treatments and finding cures for chronic, debilitating diseases, yet less than 1 % of those with a given disease participate. The biggest obstacle to patient participation is lack of awareness about clinical research studies and the role they play in the drug development process. Currently, there is no single national listing of every current clinical trial that includes procedures for enrolling and lists eligibility criteria for specific trials. Finding this information may be a hit-or-miss process, dependent on a patient's doctor, location, and access to informational resources rather than on appropriateness/eligibility criteria**.** According to CenterWatch, difficulties in patient enrollment delay 81 % of all clinical trials at least 1–6 months, with another 5 % postponed 6 months or more, and drug companies stand to lose between $600,000 and $8 million each day clinical trials delay a drug's development and launch. According to McKinsey & Co., the number of patients required to receive FDA approval has nearly doubled since the early 1990s, and more trials are needed per compound. These factors lead to longer trial timelines, which exacerbate an already burdensome issue: rising clinical costs. The average per-patient cost in the late 2000s is about $5,500 for a Phase I trial; $6,500 for a Phase II trial; and more than $7,600 for a Phase III trial. Recruiting clinical trial participants "costs more and

Table 8.1 Strategies to aid in recruitment and retention in clinical trials

Employ professional patient recruitment firms and site management organizations
Reduce the number of participants in any given trial
Outsource clinical trials to developing countries where costs are less.
Shorten trial duration by securing FDA "fast track" designation for certain drugs or treatments (primarily cancer)

consumes more time" than any other aspect of the drug development process. "With so many parties involved in clinical development—study sponsors, clinical research offices (CROs), site management organizations (SMOs), patient recruiters, investigators, and patients—it is no surprise that on average, trials last 30–42 % longer than expected" [6].

Effective recruitment offers one of the greatest opportunities to stem rising clinical costs and accelerate the trial process. Creating more efficient marketing campaigns and new ways of attracting patient populations will provide an edge as competition for patient's increases. To further improve patient recruitment and retention, trial sponsors are (as summarized in Table 8.1):

- Employing professional patient recruitment firms and site management organizations.
- Reducing the number of participants in any given trial.
- Outsourcing clinical trials to developing countries where costs are less. One-third of the participants in any given trial, however, must come from the U.S. and/or Europe to avoid skewing results because of cultural, nutritional, or economic differences of foreign trial participants from the average American trial participant.
- Shortening trial duration by securing FDA "fast track" designation for certain drugs or treatments (primarily cancer). Approval is granted before all the clinical trials are conducted, with the sponsor promising to complete the clinical trials after approval is obtained. The risk is that these "catch up" trials do not always take place.

The FDA encourages sponsors to determine why participants withdraw from studies because it could be indicative of an important safety problem. It is not helpful to record vague explanations such as "withdrew consent," "failed to return," or "lost to follow-up." Participants who leave a study because of significant safety issues should be followed closely until they are fully and permanently resolved. On December 7–8, 2012 the NHLBI convened a workshop to address successful recruitment and retention in Phase III & IV clinical trials. Three key areas were addressed: (1) public and professional awareness and acceptance of clinical trials; (2) human subject research policies, guidelines, and reimbursement; and (3) clinical trial enrollment experience and practice. Workshop participants identified several critical barriers to successful recruitment (Table 8.2). First, health care providers are gatekeepers for trial participation, yet many providers are either not aware of active trials or do not refer their patients to trials. Likewise, patients lack knowledge of clinical trials for which they may be eligible; and, often view clinical research with uncertainty or suspicion. Second, the administrative burden and regulatory requirements combined with underestimation of infrastructure, reimbursement and

Table 8.2 Some critical barriers to recruitment

Institutional policies that govern clinical investigation often create barriers for clinician investigators rather than encouragement
Many providers are either not aware of active trails or do not refer their patients to trials
Patients lack knowledge of clinical trials for which they may be eligible and often view clinical research with uncertainly or suspicion
Administrative burden and regulatory requirements combined with underestimation of infrastructure, reimbursement and resource costs associated with recruitment and retention are not fully recognized
The financial realities of funding clinical trails may not be recognized and met

From: NHLBI Workshop: Clinical Research United in Successful Enrollment (CRUiSE) [7]

Table 8.3 Approaches to barriers of successful recruitment

Engage the patients health care provider, many clinicians are either unaware of, or do not refer their patients to clinical trials
Recognize the financial realities of clinical trial funding as an ongoing challenge that includes the concept of indirect costs
Address administrative and regulatory burdens and institutional policies that create barriers for clinical investigators rather than encouraging their participation

resource costs associated with recruitment and retention present unprecedented challenges, even for experienced trialists. Third, the financial realities of funding clinical trials must be recognized and met. The unplanned expansion of time and resources often required to reach target enrollment is a major threat to the continued commitment of any sponsor. Fourth, institutional policies that govern clinical investigation often create barriers for clinician investigators rather than encouraging their participation in clinical research [7]. Many important complicating issues were identified. For example, adverse pressure on trial enrollment is an unintended consequence of the requirement to include all populations in clinical studies. Minorities and women, who are often underserved and under-insured or uninsured, are harder and more costly to recruit and retain in trials. Added costs required to enroll and retain diverse populations compete for available funds in a fiscal climate in which cost containment is encouraged. Finally, how indirect costs awarded to institutions conducting federally sponsored trials might better be used to aid clinical trial conduct was also discussed. A number of recommendations came out of this meeting and will be discussed in this chapter. To summarize, Probstfield and Frye [8] noted several barriers of successful recruitment that could be potentially addressed to improve it. One was the importance of engaging the patients gatekeeper health care provider, noting that many clinicians are either unaware of, or do not refer their patients to clinical trials. Another was the recognition of the financial realities of clinical trial funding as an ongoing challenge that includes the concept of indirect costs. The last was the unprecedented administrative and regulatory burdens and institutional policies that create barriers for clinical investigators rather than encouraging their participation (Table 8.3).

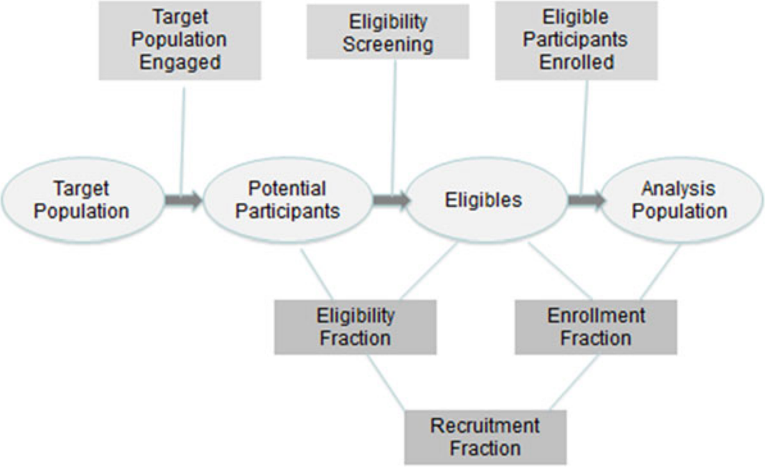


Fig. 8.1 The process of trial enrollment (From Gross et al. [10])

The recruitment process involves a number of important steps and the trial enrollment process is being increasingly addressed because of its importance to the studies ultimate generalizability [9]. An outline of the enrollment process is shown in Fig. 8.1 which also introduces a number of variables and definitions which should be considered and perhaps reported in large trials [10]. Recall that sampling (see Chap. 3) is perhaps one of the more important considerations in clinical research. Also recall, that the *target population* is the population of potentially eligible subjects, and how this is defined can have significant impact on the studies generalizability. From the target population, a smaller number are actually recruited and then enrolled (eligibility fraction and enrollment fraction). The product of these two fractions represent the proportion of potential participants who are actually enrolled in the study (recruitment fraction) [10]. An example of the use of these various fractions is taken from a study, in which we found that as defined according to standards recommended by Morton et al. [11], the response rate (percent agreeing to be interviewed among known eligible candidates contacted n = 57,253) plus an adjustment for the estimated proportion eligible among those of unknown eligibility (n = 25,581) was 44.7 % (36,983/82,834). The cooperation rate (the proportion of known eligible participants who agreed to be interviewed) was 64.6 % (36,983/57,253). This helps the reader to understand how representative the study population is. However, as Halpern has pointed out, "*although more thorough reporting would certainly help identify trials with potentially limited generalizability, it would not help clinicians apply trial results to individual patients*." [12] Halpern also points out that data on patients who chose not to participate would be important. There follows an interesting discussion of the pros and cons addressing this entire issue that is important for the interested reader. Beyond the importance of generalizability, details of the recruitment process might also demonstrate obstacles to the recruitment process.

Failures in Recruitment

Overall, clinical trial enrollment rates have dropped over time by 75 % in 2000, to 59 % in 2006, from what was initially planned; and, retention rates fell from 69 to 48 % during that same time period. This resulted in a delay of trial completion by 1 month or more in the great majority of trials. In fact, the FDA has reported that only 6 % of trials are completed on time. There are a number of reasons for failure of the recruitment process including: ethical considerations, delayed start-up, inadequate planning, insufficient effort & staff, and over-optimistic expectations. In addition recruitment to NIH studies adds an additional burden as the NIH expects adequate numbers of women, minorities and children (when appropriate) to be recruited into studies that they fund (Table 8.4).

The ethical considerations regarding recruitment are increasingly becoming an issue. Every researcher faces a critical weighing of the balance between informing patients about the benefits and risks of participating in a trial, against unacceptable encouragement to participate. IRB's are exerting increasingly more rigorous control about what is appropriate and inappropriate in this regard. This has been the subject of debate in the United Kingdom as well, and is particularly acute due to the fact that the National Health Service requires that investigators adhere to strict regulations [13]. In the UK (and to some extent in the USA), ethicists are insisting that researchers can only approach subjects who have responded positively to letters from their general practitioners or hospital clinician (the so-called 'opt in' approach). That is, under the opt-in system a subject is responsible for contacting their doctor and letting them know it is okay for a researcher to contact them. In an opt-out system, the initial letter to the patient will explain that a researcher will be contacting them unless they tell their doctor that they wish not to be contacted. Hewison and Haines have argued that the public needs to be included in the debate about what is in the subject's best interests, before an ethicist can make a unilateral decision [13]. Hewison and Haines feel that 'research ethics requirements are compromising the scientific quality of health research', and that 'opt-in systems of recruitment are likely to increase response bias and reduce response rates' [13]. There is little data on the subject of opt-in vs. opt-out systems in regards to the concerns expressed above, but the potential for bias and reduced recruitment is certainly hard to argue.

Table 8.4 Reasons for failure of the recruitment process

Ethical considerations
Delayed start-up
Inadequate planning
Over-optimistic expectations
Insufficient effort & staffing
Recruitment to NIH studies adds an additional burden as the NIH expects adequate numbers of women, minorities and children (when appropriate) to be recruited into studies that they fund

The above considerations just apply to the method of contacting potential subjects. Other issues regarding recruitment are also becoming increasingly important as more studies (particularly Industry supported studies) matriculated out of academic centers and into private offices, where the investigator and staff might not have experience in clinical research. This privatization of clinical research began in the 1990s predominantly due to the inefficiencies of working with academia, including protracted contractual and budget negotiations, bureaucratic and slow moving IRBs, and higher costs [14]. Today, only 1/3 of all industry-funded clinical trials are placed within academic centers. Now, as NIH funding is dwindling and other federal funding changes are occurring, many within academic centers are again viewing the potential of industry supported research studies.

Differences in Dealing with Clinical Trial Patients

There are differences in the handling of clinical practice patients in contrast to research subjects (although arguably this could be challenged). But at the least, research subjects are seen more frequently, have more testing performed, missed appointments result in protocol deviations, and patients lost to follow-up can erode the studies validity. In addition many research subjects are in studies not necessarily for their own health, but to help others. Thus, the expense of travel to the site, the expense of parking, less than helpful staff, and waiting to be seen may be even less tolerable than it is to clinical practice patients. Thus, the provisions for on-site child care, a single contact person, flexible appointment times, telephone and letter reminders, and the provision of study calendars with study appointment dates are important for the continuity of follow-up. In addition, at a minimum, payment for travel and time (payments to research subjects are a controversial issue) need to be considered, but not at such a high rate that the payment becomes coercive [15]. The use of financial compensation as a recruiting tool in research is quite controversial, with one major concern that such compensation will unduly influence potential subjects to enroll in a study, and perhaps even to falsify information to be eligible [16]. In addition, financial incentives would likely result in an overrepresentation of the poor in clinical trials. Also, these days, it is important that study sites maintain records of patients that might be potential candidates for trials as funding agencies are more frequently asking for documentation that there will be adequate numbers of subjects available for study. Inflating the potential for recruitment is never wise as the modified cliché goes, ‘you are only as good as your last study’. Failure to adequately recruit for a study will significantly hamper efforts to be competitive for funding for the next trial. Demonstrating to funding agencies that there is adequate staff, and facilities, and maintaining records of prior studies is also key. As Gorelick et al. point out, much has been said about the barriers in a clinical research study including mistrust, study costs, ineffective communication and lack of awareness of the importance of clinical research both by patients and health care providers. They provided a Research Triangle

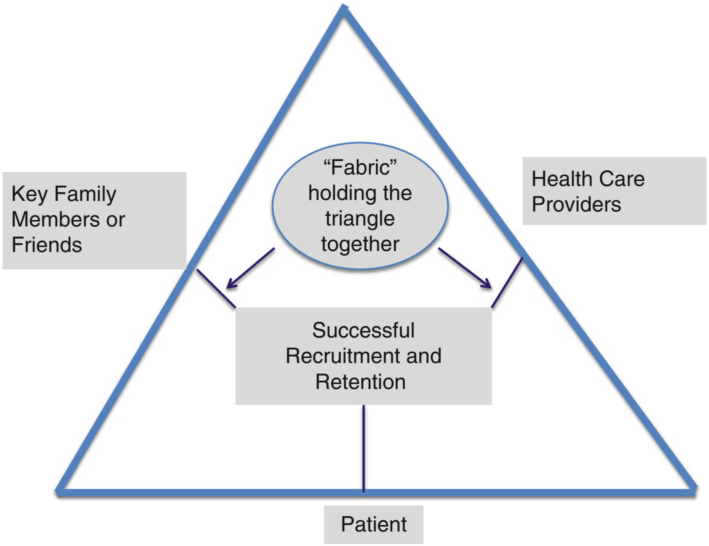

Key Family Members or Friends
"Fabric" holding the triangle together
Health Care Providers
Successful Recruitment and Retention
Patient

Fig. 8.2 The research triangle (Adapted from Gorelick et al. [17])

(Fig. 8.2) whose foundation is the patient, but where any disruption in the triangles "structure" can result in crumbling of the triangle [17].

Why People Participate in Clinical Research

There have not been many studies delving into why subjects participate in clinical research. In a study by Jenkins et al. the reasons for participating and declining to participate were evaluated (see Table 8.5) [18]. This was also evaluated by Hawkins et al. and both found that a high proportion of participants enrolled in studies to help others [19]. Hawkins et al. performed a cross sectional survey with a questionnaire mailed to 836 participants and a response rate of 31 % (n = 259). Responses were open-ended and an *a priori* category scale was used and evaluated by two research co-coordinators with a 10 % random sample assessed by a third independent party in order to determine inter-reader reliability (Table 8.6).

Few studies have attempted to objectively quantify the effects of commonly used strategies aimed at improving recruitment and retention in research studies. One that did evaluate five common strategies, assessed the effect of notifying potential participants prior to being approached; providing potential research subjects with additional information about the study; changes in the consent process; changes in the study design (such as not having a placebo arm); and; the use of incentives. The author's conclusions were that it is not possible to predict the effect of most of these approaches on recruitment [20].

Table 8.5 Reasons for participation in clinical research (2000)

Top reasons for entering	%
So others may benefit	23
Trust in the doctor	21
Cutting edge treatment	16
Top reasons for declining	
My doctor told me not to	22
Randomization worried me	20
1 wanted the doctor to treat me not a computer	18

From: Jenkins and Fallowfield [18]

Table 8.6 Reasons for participation in clinical research (1980s)

Advantages	Disadvantages
Close observation (50 %)	Inconvenience (31 %)
Self knowledge (40 %)	Side effects (10 %)
Helping others (32 %)	Symptom worsening (9 %)
Free care (25 %)	Blinding (7 %)
Improve their disease (23 %)	

From: Hawkins et al. [19]

Types of Recruitment

There are a number of additional considerations one has to make for site recruitment and retention. For example, before the study starts consideration as to how the subjects will be recruited (i.e. from a data-base, colleague referral, advertising-print, television, radio, etc.) and once the study starts there needs to be weekly targets established and reports generated, The nature of the recruitment population also needs to be considered, For example, Gilliss et al. studied the one-year attrition rates by the way in which they were recruited, and ethnicity [21]. They found that responses to and subsequent 1 year attrition rates, differed between whether the initial recruitment approach was through broadcast media, printed matter, face-to face recruitment, direct referral, and the use of the Internet; and, differed between American, African American, and Mexican American. For example, the response to broadcast media resulted in 58, 62 and 68 % being either not eligible or refusing to participate; and, attrition rates were 13, 17 and 10 % comparing American, Mexican American and African Americans respectively. In contrast, face-to-face recruitment resulted in lower refusal (21, 28, and 27 %) and attrition rates (4 %, 4, and 16 %). Remember that the effect of patient discontinuations in clinical trials is to decrease the power of the study, to decrease the studies validity and to decrease the treatment effect.

There has been increasing interest in the use of electronic medical records in the clinical trials recruitment process, partly due to the fact that clinical trials suffer from low primary care physician participation. Embi et al. performed a "before-after" analysis of a clinical trial alert system, and compared enrollment in the

Table 8.7 Features that distinguish research from clinical practice

Was the intention to produce generalizable knowledge
Was there a systematic investigation
Was there less net clinical benefit and greater risk than exists in clinical practice
Did the research introduce burdens or risks that were usually not part of the patients clinical management
Was there the use of protocols to dictate what care the patient receives

12 months before and 4 months after its implementation [22]. During the 12- month "control period" for a specific diabetes mellitus trial, they found an enrollment rate of 2.9 participants/month. During the 4 months following the implementation of their EMR alert system this rose to 6.0/month. However, it is apparent that better systems need to be developed. For example the majority of the information in EMRs is contained in clinical notes (i.e. in free text format), and there is variability in data sources capturing the needed information for recruiting specific patients into a trial. There is also the issue of incentivizing practitioner involvement in clinical trials. One approach is to design a pragmatic trial that can mimic, and therefore be easily incorporated into clinical practice. In the Hastings Center Report [23], five features were listed that were used to distinguish research from clinical practice as follows: was the intention to produce generalizable knowledge, was there a systematic investigation, was there less net clinical benefit and greater risk than exists in clinical practice, did the research introduce burdens or risks that were usually not part of the patients clinical management, and was there the use of protocols to dictate what care the patient receives (Table 8.7).

Minority Recruitment

The inclusion of diverse populations in clinical trials is central to generating generalizable findings and understanding health disparities; and, this has generated increased interest in enrolling minorities into clinical research trails. In fact, despite the recognition that minority groups are disproportionally affected by, for example, cardiovascular and renal disease, ethnicity-specific analyses have in the past, been generally inadequate for determining subgroup effects.

In 1993, the National Institutes of Health Revitalization Act mandated minority inclusion in RCTs, and defined underrepresented minorities as African Americans, Latinos, and American Indians. Subsequently, review criteria have formally required minority recruitment plans or scientific justification for their exclusion. Yancey et al. [24], evaluated the literature on minority recruitment and retention and identified ten major themes or factors that emerged as influencing minority recruitment. Further, they noted that if addressed appropriately these ten themes facilitated recruitment: attitudes towards perceptions of the scientific and medical community; sampling approach; study design; disease specific knowledge and perceptions of

Table 8.8 Major factors that influence minority recruitment

Attitudes towards perceptions of the scientific and medical community
Facilitated recruitment
Sampling approach
Study design
Disease specific knowledge and perceptions of prospective participants
prospective participants psychosocial issues
Community involvement
Study incentives and logistics
Sociodemographic characteristics of prospective participants
Participant beliefs such as religiosity; and cultural adaptations or targeting

From: Yancy et al. [24]

prospective participants; prospective participants psychosocial issues; community involvement; study incentives and logistics; sociodemographic characteristics of prospective participants; participant beliefs such as religiosity; and cultural adaptations or targeting (Table 8.8). In general, most of the barriers to minority participation were similar for non-minorities except for the greater mistrust by African Americans toward participation (particularly into interventional trials), likely as a result of past problems such as the Tuskegee Syphilis Study [25]. Some of the authors conclusions based upon their review of the literature included: mass mailing is effective; population-based sampling is unlikely to produce sufficient numbers of ethnic minorities; community involvement is critical; and, survey response rates are likely to be improved by telephone follow-up.

Despite the aforementioned studies, the aggregate literature is conflicting, and the majority of it is broad and descriptive. Indeed, despite the generally held belief that minority groups are less willing to participate in research (due to mistrust or perceived if not real discrimination) literature on this subject is by no means conclusive. One group evaluated the attitudes among African American candidates who were recruited to the African American Study of Kidney Disease and Hypertension (AASK Trial) [26]. They compared candidates to the study who chose to participate and those who refused enrollment and found that the most significant predictors of enrollment was the perceived impact of study participation on their health status, while mistrust and health-related factors did not emerge as barriers to participation. In the study by Martin et al. although they did find increased difficulty in enrolling adequate numbers of females and elderly (age >65 year) there was a lack of effect of race on willingness to participate in RCTs [27]. When it comes to other minorities, there is even less data available. For instance, there is evidence of under-representation of women in study populations, and data suggests that women are less willing to participate in research, but the reasons are unclear and apparently complex. Ding et al. did attempt to evaluate how willingness to participate affected participation in men vs. women [28]. They found that women showed less distrust of medical researchers, but perceived a greater risk of harm compared to men. Overall the RR of willingness to participate was 1.15 (CI 1.02–1.31) in men compared to women. Clearly more answers are needed in this area. More recently, Martin et al. attempted to objectively quantify patient and trial-specific factors with participation in cardiovascular RCTs [27]. They studied a consecutive sample of patients who were screened and potentially eligible for participation into a RCT, with a main outcome measure of not participating in a RCT. They found that trial-specific factors were more strongly associated with non-participation than patient-specific factors. Two specific trial-related factors associated with non-participation were intensive testing and participation >6 months, while patient specific factors were age >65 year, female sex, and residence location. Although other trial specific factors were also different between participants and non-participants (Industry sponsored trials, study size, trial compensation <$20, and outpatient recruitment) these were not evident upon multivariable analysis (Fig. 8.3). Another interesting aspect of this study was the discordant views that the trial team had vs. the patient's reasons for non-participation. The trial team listed things like potential non adherence, altered mental status, clinical instability and substance abuse, while patients talked about travel barriers and being too busy (again, payments for participation and fear of the drug being used in the trial were very infrequent reasons for non-participation).

Although not included in the NIH definition of "minorities" it has been recognized that women and the elderly have also not been adequately represented in past clinical trials. For instance, Cherubuni et al. investigated the extent of exclusion of older individuals in clinical trials of heart failure, a condition in which the elderly are inordinately represented [29]. They found that among 251 trials, 64 (25.5 %) excluded patients at an arbitrary upper age limit, and that such exclusions were more common amongst European trials than those conducted in the US. Overall, they concluded that "109 trials (43.4 %) had 1 or more unjustified exclusion criterion that could limit the inclusion of older individuals". In a review article by Heiat et al. they noted that there was no significant improvement in elderly recruitment into heart failure trials among publications from the 1980 and early 1990s compared to those published later [30].

In 2011 Burke et al. assessed (1) the reporting of race, ethnicity, and gender in trial publications and (2) the enrollment of women and minorities in NINDS-funded phase III clinical trials with published results [31]. Representation of women and minorities in stroke related trials was assessed separately to allow for comparison with the relatively well-described epidemiology of stroke. Between 1985 and 2008 56 trials reported enrollment by gender, race, and ethnicity were identified, and the percent of African Americans, Hispanic Americans, and women enrolled in the trials was calculated, for those trials with available data. African Americans constituted 19.8 % of the enrollees in trials with available data and enrollment increasing over time (11.6 % period 1; 30.7 % period 2, p _ 0.001). Hispanic Americans constituted 5.8 % of subjects in trials with available data and enrollment decreased over time (7.4 % period 1; 5.0 % period 2, p _ 0.001). The

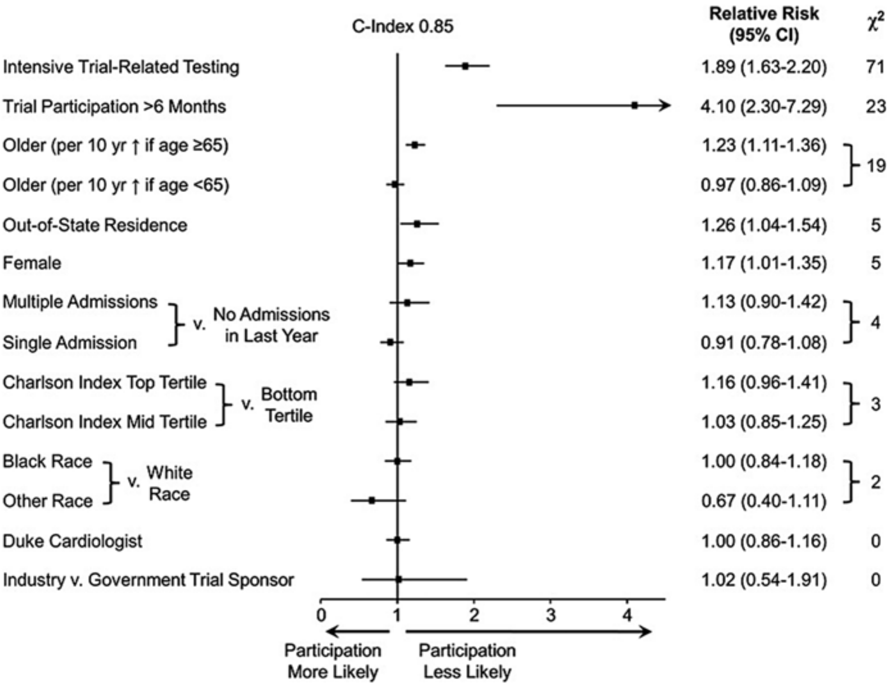


Fig. 8.3 Multivariable analysis of factors associated with not participating in a cardiovascular randomized clinical trial
The multivariable model included all of the factors in the Forest plot. Factors are ordered from top to bottom by highest to lowest chi-square. Relative risks and 95 % confidence intervals (CI) are presented. The analysis included the primary analysis sample of 655 subjects without a language barrier. Intensive trial-related testing was defined as cardiac magnetic resonance imaging, retinal examinations requiring extended study visits, or peak oxygen uptake measurement during exercise. Chronic comorbidity burden was quantified by the Charlson Index. Hospital admissions were for any cause within the past year. Duke Cardiologist was defined as having an outpatient encounter within the past year with a Duke University Medical Center-affiliated cardiologist. Martin et al. [29]

authors concluded that "*in the 15 years since implementation of the NIH Revitalization Act, representation of women and African Americans in clinical trials has improved significantly*". However, they cautioned that despite clear progress, underrepresentation persists in a number of specific disease states, including stroke and other neurologic diseases [31].

HIPPA

A final word about recruitment relates to HIPAA (the Health Insurance Portability and Accountability Act). Issued in 1996, the impetus of HIPPA was to protect patient privacy. However, many have redefined HIPAA as *'How Is it Possible to*

Accomplish Anything'. As it applies to research subjects it is particularly confusing. The term protected health information (PHI) includes what physicians and other health care professionals typically regard as a patient's personal health information. PHI also includes identifiable health information about subjects of clinical research gathered by a researcher. Irrespective of HIPAA, the safeguarding of a patient's personal medical records should go without saying; and, failure of this aforementioned safeguarding has resulted in problems for some researchers. As it affects patient recruitment, however, HIPAA is problematic in that the researcher's ability to contact patients for a research study, particularly patients of another health care provider, becomes more problematic. In addition, in clinical research, the investigator is often in a dual role as it regards a patient—that of a treating physician and that of a researcher. Long-standing ethical rules apply to researchers, but in regard to HIPAA, a researcher is not a 'covered entity' (defined as belonging to a health plan, health care clearinghouse, or health care provider that transmits health information electronically). However, complicating the issue is when the researcher is also a health care provider, or employees or other workforce members are a covered entity. The role and scope of HIPAA, as it applies to clinical research is beyond the intention (or comprehension) of this author and, therefore, will not be further discussed.

Summary

During my over 35 years of clinical research experience I have developed a number of strategies aimed at retaining participants, and some examples are outlined below.

- *A participant in a 4 year outcome trial discontinued study drug over 1 year ago (during the second year of the trial) due to vague complaints of weakness and fatigue, however, the participant did agree to continue to attend study visits. At one of the follow up visits, we asked the participant if they would be willing to try the study drug again, and in so doing were able to re-establish the participant in the trial.* Recall that based upon the intention-to-treat principle (see Chap. 3) they would have been counted as having received their assigned therapy anyway, and in terms of the outcome it is still better that they received the therapy for 3 of the 4 years, than for less than that.
- *Another participant reported a loss of interest in the study and stopped his study drug. Upon questioning it was determined that he had read newspaper articles about recent studies involved with the study drug you are testing, and felt there is nothing to gain from continuing in the study. We explained how this study differs from those reported in the newspaper, using a fact based approach, and the subject was willing to participate once again.*
- *A participant following up on the advice of his primary care doctor (PCP) decided he would like to know what study drug he was receiving when the PCP noted a BP of 150/90 mmHg. Further, the PCP had convinced the patient to discontinue blinded study therapy. You receive a call from the patient stating they no longer wish to participate in the study.* One way of preventing this in the first place is to

involve the patients PCP from the beginning. However, in this case, the patient had transferred to a new PCP and had not informed us. As a result, we called the PCP and communicated the importance of the study and assured the PCP that better BP control is expected and that we would be carefully monitoring his BP.

In summary, a frank open discussion with the patient as to what happened and why he/she wants to discontinue is important, as is preserving rapport with the patient and their PCP, that is a key to subject retention. It is also critical that the principal investigator (PI) maintain frequent contact (and thereby solidify rapport) with the patient, given that in many studies the study coordinator and not the PI may see the patient on most occasions. I remember asking one study coordinator if they knew the definition of PI and the immediate response was 'yes-practically invisible!'

References

1. Kim ES, Carrigan TP, Menon V. International participation in cardiovascular randomized controlled trials sponsored by the National Heart, Lung, and Blood Institute. J Am Coll Cardiol. 2011;58:671–6. doi:10.1016/j.jacc.2011.01.066.
2. Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, et al. Recruitment to randomised trials: strategies for trial enrollment and participation study. The STEPS study. Health Technol Assess. 2007;11(iii):ix–105.
3. Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. BMJ. 1984;289:1281–4.
4. Easterbrook PJ, Matthews DR. Fate of research studies. J R Soc Med. 1992;85:71–6. PMC1294885.
5. Pound ET. A case study in how not to conduct a clinical trial. USA Today. 2000.
6. Cutting edge information White Paper: accelerating clinical trials: budgets, patient recruitment and productivity. Accessed at www.cuttingedgeinfo.com/research/clinical-development. Accessed 15 July 2013.
7. NHLBI workshop: Clinical Research United in Successful Enrollment (CRUiSE). Accessed at www.nhlbi.nih.gov/meetings/workshops/cruise.htm. Accessed 2 Dec 2013.
8. Probstfield JL, Frye RL. Strategies for recruitment and retention of participants in clinical trials. JAMA. 2011;306:1798–9. doi:10.1001/jama.2011.1544.
9. Wright JR, Bouma S, Dayes I, Sussman J, Simunovic MR, Levine MN, et al. The importance of reporting patient recruitment details in phase III trials. J Clin Oncol. 2006;24:843–5.
10. Gross CP, Mallory R, Heiat A, Krumholz HM. Reporting the recruitment process in clinical trials: who are these patients and how did they get there? Ann Intern Med. 2002;137:10–6.
11. Morton LM, Cahill J, Hartge P. Reporting participation in epidemiologic studies: a survey of practice. Am J Epidemiol. 2006;163:197–203.
12. Halpern SD. Reporting enrollment in clinical trials. Ann Intern Med. 2002;137:1007–8; author reply –8.
13. Hewison J, Haines A. Overcoming barriers to recruitment in health research. BMJ. 2006; 333:300–2.
14. Getz K. Industry trials poised to win back academia after parting ways in the late 90s. Appl Clin Trials. 2007. Available at: www.appliedclinicaltrialsonline.com/appliedclinicaltrials/article/articleDetail.jsp?id=416536
15. Giuffrida A, Torgerson DJ. Should we pay the patient? Review of financial incentives to enhance patient compliance. BMJ. 1997;315:703–7.

16. Dunn LB, Gordon NE. Improving informed consent and enhancing recruitment for research by understanding economic behavior. JAMA. 2005;293:609–12.
17. Gorelick PB, Harris Y, Burnett B, Bonecutter FJ. The recruitment triangle: reasons why African Americans enroll, refuse to enroll, or voluntarily withdraw from a clinical trial. An interim report from the African-American Antiplatelet Stroke Prevention Study (AAASPS). J Natl Med Assoc. 1998;90:141.
18. Jenkins V, Fallowfield L. Reasons for accepting or declining to participate in randomized clinical trials for cancer therapy. Br J Cancer. 2000;82:1783–8.
19. Hawkins C, West T, Ferzola N, Preismeyer C, Arnett D, Glasser S. Why do patients participate in clinical research? In: Associates of clinical pharmacology 1993 annual meeting; 1993.
20. Mapstone J, Elbourne DR, Roberts I. Strategies to improve recruitment in research studies. Cochrane Database Syst Rev. 2007;18, MR000013.
21. Gilliss C, Lee K, Gutierrez Y, Taylor D, Beyene Y, Neuhaus J, et al. Recruitment and retention of healthy minority women into community-based longitudinal research. J Womens Health Gend Based Med. 2001;10:77–85.
22. Embi PJ, Jain A, Clark J, Bizjack S, Hornung R, Harris CM. Effect of a clinical trial alert system on physician participation in trial recruitment. Arch Intern Med. 2005;165:2272–7.
23. Kass NE, Sugarman J, Faden R, Schoch-Spana M. Trust, the fragile foundation of contemporary biomedical research. Hastings Cent Rep. 1996;26:25–9.
24. Yancy AK, Ortega AN, Kumanyika SK. Effective recruitment and retention of minority research participants. Annu Rev Public Health. 2006;27:1–28.
25. Tuskegee Syphilis Study. Accessed at www.tuskegee.edu/Global/Story.asp?s=1207598
26. Gadegbeku CA, Stillman PK, Huffman MD, Jackson JS, Kusek JW, Jamerson KA. Factors associated with enrollment of African Americans into a clinical trial: results from the African American study of kidney disease and hypertension. Contemp Clin Trials. 2008;29:837–42.
27. Martin SS, Ou FS, Newby LK, Sutton V, Adams P, Felker GM, et al. Patient- and trial-specific barriers to participation in cardiovascular randomized clinical trials. J Am Coll Cardiol. 2013;61:762–9. doi:10.1016/j.jacc.2012.10.046.
28. Ding EL, Powe NR, Manson JE, Sherber NS, Braunstein JB. Sex differences in perceived risks, distrust, and willingness to participate in clinical trials: a randomized study of cardiovascular prevention trials. Arch Intern Med. 2007;167:905–12.
29. Cherubini A, Oristrell J, Pla X, Ruggiero C, Ferretti R, Diestre G, et al. The persistent exclusion of older patients from ongoing clinical trials regarding heart failure. Arch Intern Med. 2011;171:550–6. doi:10.1001/archinternmed.2011.31.
30. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. Arch Intern Med. 2002;162:1682–8.
31. Burke JF, Brown DL, Lisabeth LD, Sanchez BN, Morgenstern LB. Enrollment of women and minorities in NINDS trials. Neurology. 2011;76:354–60. PMC3034419.

Chapter 9
Data Safety and Monitoring Boards (DSMBs)

Stephen P. Glasser and O. Dale Williams

Abstract Data Safety and Monitoring Boards were introduced as a mechanism for monitoring interim data in clinical trials as a way to ensure the safety of participating subjects. Procedures for and experience with DSMBs has expanded considerably over recent years and they are now required by the NIH for almost any interventional and for some observational trials. A DSMB's primary role is to evaluate adverse events and to determine the relationship of the adverse event to the therapy (or device). Interim analyses and early termination of studies are two aspects of DSMBs that are particularly difficult challenges. This chapter will discuss the role of DSMBs and address the aforementioned issues.

Keywords DSMB • Interim analysis • Early study termination • Multiple comparisons • OSMB

Data Safety and Monitoring Boards (DSMBs), which have various names including Data Safety and Monitoring Committees and Data Monitoring Committees, were born in 1967, a result of a National Institute of Health (NIH) sponsored task force report known as the Greenberg Report [1]. Initially the responsibilities now assigned to a DSMB were a component of those of a Policy Advisory Board. From this emerged a subcommittee to focus on monitoring clinical trial safety and efficacy. More specifically, the DSMB was introduced as a mechanism for monitoring interim

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

O.D. Williams, Ph.D.
Department of Biostatistics, Robert Stempel College of Public Health and Social Work,
Florida International University, Miami, FL, USA

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_9,

data in clinical trials as a way to ensure the safety of participating subjects. Procedures for and experience with DSMBs has expanded considerably over recent years, and several key publications relevant to their operations are now available [2–4]. In general, the NIH now requires DSMB's for all clinical trials (including some phase 1 and 2 trials), and recently added device trials to this mandate [5]. The NIH website states *"Applications that include clinical trials must include a general description of the data and safety monitoring plan. The description of the data and safety monitoring plan in competing applications will be reviewed by the Scientific Review Group (SRG). A general description of a monitoring plan establishes the overall framework for data and safety monitoring. It must describe the entity that will be responsible for monitoring how adverse events will be reported to the IRB and the NIH and, when appropriate, how the NIH Guidelines and FDA regulations for INDs and IDEs will be satisfied."*

"A detailed monitoring plan must be included as part of the research protocol, be submitted to the local IRB, and be reviewed and approved by the NIH awarding IC prior to the accrual of human subjects. The awarding IC may specify the reporting requirements for adverse events, which are in addition to the annual report to the IRB. The clinical trial monitoring function is above and beyond that traditionally provided by IRBs; however, the IRB must be cognizant of the procedures used by clinical trial monitoring entities and the monitor must provide periodic reports to investigators for transmittal to the local IRB." "NIH specifically requires the establishment of DSMBs for multi-site clinical trials involving interventions that entail potential risk to the participants, and generally for Phase III clinical trials. Although Phase I and Phase II clinical trials also may use DSMBs, smaller clinical trials may not require this oversight format, and alternative monitoring plans may be appropriate." [6] DSMB's are now an established interface between good science and good social values. For example, the National Heart Lung and Blood Institute (NHLBI) at the NIH [7] requires the following:

- For Phase III clinical trials, a Data and Safety Monitoring Board (DSMB) is required. This can be a DSMB convened by the NHLBI, or by the local institution, depending on the study, the level of risk and the funding mechanism.
- For a Phase II trial, a DSMB may be established depending on the study, but in most cases a DSMB appointed by the funded institution may suffice.
- For a Phase I trial, monitoring by the PI and the local IRB usually suffices. However, a novel drug, device or therapy with a high or unknown safety profile may require a DSMB.
- For an Observational Study, a Monitoring Board (OSMB) may be established for large or complex observational studies. This would be determined on a case-by-case basis by NHLBI.

More specifically, as an investigator, the questions one should ask themselves in regards to whether a DSMB is likely to be needed are outlined in Table 9.1, and revolve around the study size, number of study sites, expectation of safety issues, and whether the outcomes entail mortality and/or significant morbidity.

Table 9.1 Indications that a DSMB is warranted

Large study population and/or multiple sites
Trial is intended for proving efficacy/effectiveness and/or safety of an intervention
Is there the potential for significant toxicity
Is mortality or another major endpoint the trial outcome
Is early stopping a possibility

Table 9.2 The roles of a DSMB

Identify slow rates of patient/subject accrual
Identify high rates of ineligibility after randomization
Identify protocol violations
Identify high rates of dropouts
Ensure the overall credibility of the study
Ensure the validity of study results
Protect trial participant safety

The NHLBI also requires that each DSMB operate under an approved Charter [3], with the expectation that this Charter will delineate that the primary function of the DSMB will be to ensure patient safety, as well as to ensure that patients are adequately informed of the risk in study participation. The DSMB Charter requires a formal manual of operations (MOOP) and the DSMB and sponsor must agree on all the terms set forth in the MOOP (this is sometimes referred to a Data Safety and Monitoring Plan – DSMP). This includes such things as the DSMBs responsibility, its membership, meeting format and frequency, specifics about the decision making process, report preparation, whether the DSMB will be blinded or not to the treatment arms, and the statistical guidelines that will be utilized by the DSMB to determine whether early termination of the study is warranted. In addition, DSMBs assure that the rate of enrollment is sufficient to achieve adequate numbers of outcomes, develop guidelines for early study termination, and to evaluate the overall quality of the study to include accuracy, timeliness, data flow, etc.

The DSMB is charged with assessing the progress of clinical trials and to recommend whether the trail should continue, be modified, or discontinued. More specifically, the DSMB approves the protocol, has face-to-face meetings usually every 6 months (these are supplemented with conference calls), they may have subgroup meetings for special topics, and are on call for crises; and, DSMBs review interim analyses (generally required for NIH studies). An interim analysis is one performed prior to study completion. The role of the DSMB is outlined in Table 9.2 and includes identifying slow rates of accrual and high rates of ineligibility, identify protocol violations, dropout rates validity of the study and study safety.

Members of DSMBs are to be expert in areas relevant to the trial, be approved by the sponsor, and have no conflicts of interest relative to the study to be monitored. The members should not be affiliated with the sponsor, and should be independent from any direct involvement in the performance of the clinical trial. Members of the DSMB tend to include clinical experts, statisticians, ethicists, and community representatives. Thus, the DSMB's overarching objectives are to ensure the safety of participants, oversee the validity of data, and to provide a mechanism for the early termination of studies.

Early Study Termination

As mentioned prior, the DSMB's primary role is to evaluate adverse events and to determine the relationship of the adverse event to the therapy (or device). As the DSMB periodically reviews study results, evaluates the treatments for excess adverse effects, determines whether basic trial assumptions remain valid, and judges whether the overall integrity of the study remains acceptable, it ultimately makes recommendations to the funding agency. For NHLBI sponsored studies, this recommendation goes directly to the Institute Director, who has the responsibility to accept, reject, or modify the DSMB recommendations.

The issue of terminating a study early or of altering the course of its conduct is a critically important decision. On the surface, when to terminate a study can be obvious such as when the benefit of intervention is so clear that continuing the study would be inappropriate, or conversely when harm is clearly evident. That is, a study should be stopped early if bad is happening, good is happening, or nothing is happening (that is, the prospects are poor that if the study continues there will be benefit). Finally, the DSMB can recommend early termination if there are forces external to the study that warrant its early discontinuation (e.g. a new life saving modality is approved during the course of the study that might benefit study participants). More frequently, however, it is difficult to sort out this balance of risk vs. benefit, and judgment is necessary. As Williams so aptly put it 'stopping too early is to soon, and too late is not soon enough' i.e. no one is going to be happy in either case [8]. That is, stopping a trail to early leads to results that may not be judged to be convincing, might impact other ongoing studies, or that endpoints not yet adjudicated may affect the results of the study. Finally, the DSMB must be concerned with the potential for operational chaos that may ensue, and unnecessary costs may be realized when a study is terminated ahead of schedule; however, stopping a trial to late may be harmful to patients. In addition one may keep society waiting for potentially beneficial therapy.

Another dilemma faced by early stopping is if the trial is in its beginning phases, and an excess of adverse events, for example, is already suggested. The DSMB is then faced with the question of whether this observation is just a 'blip' which will not be evident for the rest of the trial and stopping at this point would hamper if not cause cessation of a drugs development. If, on the other hand it is in the middle of the trial and efficacy, for example, is not yet fully demonstrated, the question faced by the DSMB is whether there can still be a turnaround such that the results will show benefit. Finally, if it is late in a trial, and there has been no severe harm demonstrated, but apparent efficacy is minimal, the question is whether it is worth the cost and confusion to stop the trial before completion, when it will be ending shortly anyway. In the final analysis, it is generally agreed that the recommendation to modify or terminate a trial should not solely be based upon statistical grounds. Rather, 'no statistical decision, rule, or procedure can take the place of the well reasoned consideration of all aspects of the data by a group of concerned, competent, and experienced persons with a wide range of scientific backgrounds and points of view' [9]. An example of this dilemma is represented by the AIM-HIGH study, whose primary end point was the first event of a composite of CHD death, nonfatal MI, ischemic stroke, hospitalization for ACS, or symptom-driven coronary or cerebral revascularization. At an interim analysis the DSMB found that there was a similar outcome in the two groups, occurring in 282 patients (16.4 %) in the niacin group vs. 274 patients (16.2 %) on placebo. There was also no difference in two secondary composite end points. The trial was stopped early in 2012 after a mean follow-up of 3 years. The statement from the National Heart Lung and Blood Institute at the time said the trial had been stopped because niacin was showing no additional benefits over placebo and there was also a small, unexplained increase in ischemic stroke in the niacin group.

On the stroke issue, the AIM-HIGH investigators reported that ischemic stroke occurred as the first event in 27 niacin patients (1.6 %) vs. 15 placebo patients (0.9 %). However, eight of the strokes in the niacin group occurred between 2 months and 4 years after discontinuation of niacin treatment. "When all the patients with ischemic stroke were considered, rather than just those in whom the stroke was the first study event, a nonsignificant higher trend persisted in the niacin group." One medical article questioned whether this slight increased stroke risk (p value was 0.11) was "a causal association or the play of chance"? One medical expert was quoted as saying *"I do not agree with the decision to stop this trial. It was completely inappropriate. The NIH sponsors saw a weak signal of stroke and panicked, and when all the data have come in, this doesn't appear to be an issue. Now we have lost the opportunity to properly answer the very important question of whether niacin adds any benefit in this population with low LDL levels"*.

Interim Analysis

Interim analyses may occur under two general circumstances; based on accrual – e.g. one interim analysis after half of the patients have been evaluated for efficacy (this to some degree depends on the observation time for efficacy), or based on time – e.g. annual reviews. Often the information fraction (the number of events that have occurred compared to those expected) provides a frame of reference [10].

Stopping rules for studies, as mentioned before, are dependent upon both known science and judgment. For example, in a superiority trial if one treatment arm is demonstrating ‘unequivocal’ benefit over another, study termination can be recommended. However there are problems with this approach. For example one of the study arms may show superiority at the end of year 1, but may then lose any advantage over the ensuing time period of the study. A way of dealing with this at the time of the interim analysis is to assess futility. That is, given the recruitment goals, if at the time of the study, an interim analysis suggests that there is no demonstrable difference in the treatment arms, and one can show that it would be unlikely (futile) that with the remaining patients a statistically significant difference is likely to occur, the study can be stopped [11].

Finally, an issue with interim analysis is the multiple comparisons problem (see Chap. 3). In other words, with each interim analysis, sometimes called a ‘look,’ one ‘spends’ some of the overall alpha level. Alpha is, of course, the overall significance level (usually <0.05). Statisticians have developed various rules to deal with the multiple comparison problem that arises with interim data analysis. One approach is to stop trials early only when there is overwhelming evidence of efficacy. Peto has suggested that overwhelming evidence is when $p < 0.001$ for a test that focuses on the primary outcome [12]. Perhaps the simplest method to understand is the Bonferroini adjustment, which divides the overall alpha level by the number of tests to be conducted to obtain the alpha level to use for each interim test. As discussed in Hulley et al. [13] that means that if five tests are done and the overall alpha is 0.05, then for statistical significance for stopping, a $p < 0.01$ or less, for each individual test is needed. This latter approach is typically conservative in that the actual overall alpha level may be well below 0.05.

There often are compelling reasons to make it more difficult to cross a stopping boundary early rather than later in the study. Hence, another approach is to have a stopping boundary that changes as the trial moves closer to its predetermined end, with higher boundaries earlier and lower ones later. The rationale is that early in the study, the number of endpoint events is typically quite small and thus trends are subject to high variability. This makes it more likely that there is a more extreme difference between the treatment arms early that will settle down later. Also, as the end of the trial nears, a less stringent p value is required to indicate significance, since the results are less likely to change (there will be fewer additional patients added to the trial compared to earlier in its conduct) [10].

The three most commonly used methods for setting boundaries, sometime referred to as group sequential boundaries, as a frame of reference for early termination decisions are: the Haybittle-Peto [12, 14], Pocock [15], and Obrien-Fleming [16] methods. The Haybittle-Peto and Pocock methods do not provide higher boundaries early in the study, whereas, the Obrien-Fleming, and the Lan-Demets [10] modification do. Figure 9.1 shows how these compare to each other for situations whereby five “looks” are expected for the trial [17, 18]. Thus, interim safety reports pose well recognized statistical problems related to the multiplicity of statistical tests conducted on the accumulating set of data. The basic problem is well known and is referred to as “sampling to a foregone conclusion” [17], or the problem of repeated significance tests [19, 20].

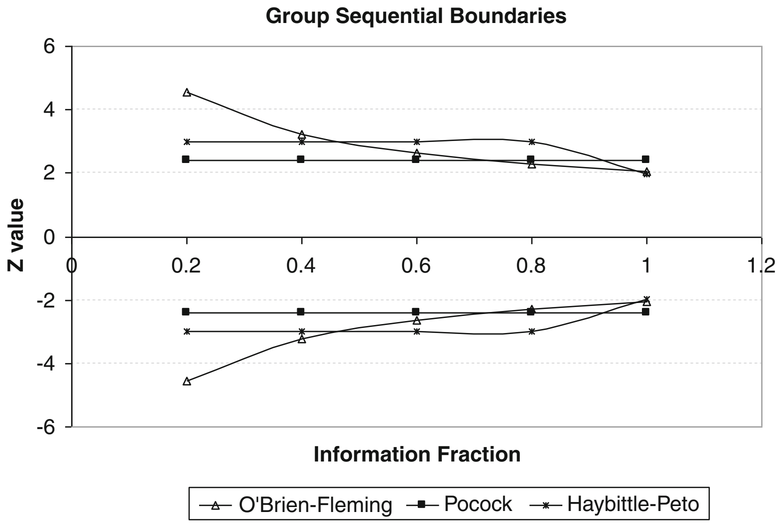


Fig. 9.1 A diagram comparing the different statistical approaches for interim analyses

Califf et al. outlined the questions that should be asked by a DSMB before altering a trial [21]. Califf et al. also point out that statistical rules are not absolute but provide guidance only. Some additional issues discussed in their review include the role of the DSMB in event-driven (i.e. the trial continues until the pre-specified number of events has been accrued) vs. fixed-sample, fixed-duration trials; how the DSMB should approach equivalence vs. noninferiority trials; the role of a Bayesian approach to DSMB concerns; the use of asymmetric vs. symmetric boundaries (the threshold for declaring that a trial should be stopped, should be less stringent for safety issues than it is when a therapy shows a positive result); and, perhaps most importantly, the overall philosophy of early stopping-that is, where does the committees primary ethical obligation lie, and what degree of certainty is required before a trial can be altered [19].

The disadvantages of stopping a trial early are numerous. These include the fact that the trial might have been terminated on a random 'high'; the reduction in the credibility of the trial when the number of patients studied will have been less than planned; and, the greater imprecision regarding the outcome of interest as the smaller sample size will have resulted in wider confidence limits. Montori et al. performed a systematic review of randomized trials stopped early as a result of their demonstrating benefit at the time of an interim analysis [22]. They noted that 'taking the point estimate of the treatment effect at face value will be misleading if the decision to stop the trial resulted from catching the apparent benefit of treatment at a 'random high'. When this latter situation occurs, data from future trials will yield a more conservative estimate of the treatment effect, the so called "regression to the truth effect". Montori's findings suggested that there were an increasing number of RCTs reported to have stopped early for benefit; and, that the early stopping

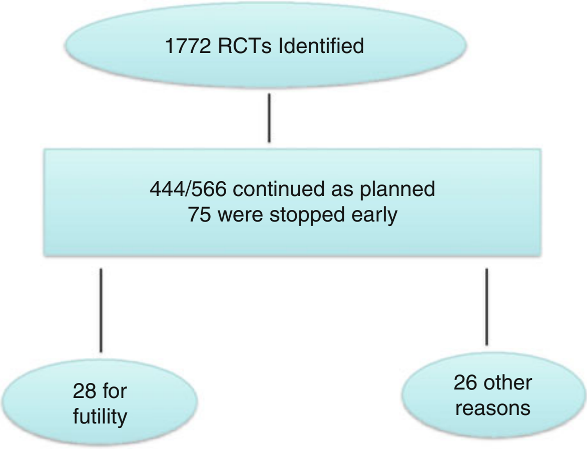


Fig. 9.2 A diagram outlining how often and for what reasons interim-analyses resulted in early stopping of a clinical trial (From: Tharmananathan et al. [24])

occurred with (on average) 64 % of the planned sample having been entered. More importantly, they concluded that information regarding the decision to stop early was inadequately reported, and overall such studies demonstrated an implausibly large treatment effect and they then suggest that the results of such studies should be viewed with skepticism [20]. One example of early stopping for harm was the ILLUMINATE trial which was terminated early by the DSMB because the trial drug, Pfizer's *torcetrapib*, had more events than placebo [23]. The questions addressed but not able to be answered were: Why this occurred? Was it the drug itself or the dose of the drug? What was the mechanism of adverse events/, etc.

Finally, as discussed in Chap. 3 the duration of the clinical trial can be an important consideration in the DSMB deliberations. Some studies may show early lack of benefit and have a delayed beneficial effect. The DSMB should carefully follow the curves elucidating the study endpoints in order to identify the potential for a delayed effect. Thus, the DSMB might not be only involved in early stopping, but might suggest a longer duration of the RCT than originally planned.

Finally, most journals endorse the CONSORT Statement (See Chapter 19) which states "when applicable, explanation of any interim analyses and stopping guidelines" should be included. As reported by Tharmananathan et al. [24], this does not happen as often as it should. Figure 9.2 is a diagram outlining how often and for what reasons interim analyses resulted in earl stopping [24].

Observational Study and Monitoring Boards (OSMBs)

OSMBs are a more recent development and are not as often necessary as they are with interventional trials [25]. Thus, a main question is when should an OSMB be established? It is the policy of the NHLBI to establish OSMBs for Institute-sponsored observational studies and registries when an independent group is needed to evaluate the data on an ongoing basis to ensure participant safety and/or study integrity. The decision to establish an OSMB is made by the relevant Division Director with the concurrence of the Director, NHLBI. As a general rule, the NHLBI appoints OSMBs for:

- all large, long-term Institute-initiated and selected investigator-initiated observational studies, whether multiple or single center in nature; and
- selected smaller Institute-initiated and selected investigator-initiated observational studies or registries to help assure the integrity of the study by closely monitoring data acquisition for comprehensiveness, accuracy, and timeliness; and monitoring other concerns such as participant confidentiality.

The role of the OSMBs is similar to that of the DSMB, that is, to monitor study progress and to make recommendations regarding appropriate protocol and operational changes. They also address safety issues such as those involving radiation exposure or other possible risks associated procedures or measurements that are study components. Decisions to modify the protocol or change study operations in a major way may have substantial effects upon the ultimate interpretation of the study or affect the study's funding. Thus, OSMBs play an essential role in assuring quality research. The principal role of the OSMB is to monitor regularly the data from the observational study, review and assess the performance of its operations, and make recommendations, as appropriate with respect to:

- the performance of individual centers (including possible recommendations on actions to be taken regarding any center that performs unsatisfactorily);
- issues related to participant safety and informed consent, including notification of and referral for abnormal findings;
- adequacy of study progress in terms of recruitment, quality control, data analysis and publications;
- issues pertaining to participant burden;
- impact of proposed ancillary studies and substudies on participant burden and overall achievement of the main study goals; and
- overall scientific directions of the study.

Thus, the OSMB must provide a multidisciplinary and objective perspective, with expert attention to all of these factors during the course of the study, and considerable judgment.

The responsibilities of the OSMBs are summarized in a document that can be found on the NHLBI web site [24].

References

1. Organization, review, and administration of cooperative studies (Greenberg Report): a report from Heart Special Project Committee to the National Advisory Heart Council. Control Clin Trials. 1967;9:137–48.
2. DeMets D, Furberg C, Friedman L. Data monitoring in clinical trials. A case studies approach. New York: Springer; 2006.

3. Ellenberg SS, Fleming TR. Data monitoring committees in clinical trials. A practical perspective. West Sussex: Wiley; 2002.
4. Friedman L, Furberg C, DeMets D. Fundamentals of clinical trials. 3rd ed. New York: Springer; 1998.
5. Connolly SJ. Use and misuse of surrogate outcomes in arrhythmia trials. Circulation. 2006;113:764–6.
6. NIH policy and guidelines on the inclusion of women and minorities as subjects in clinical research. http://grants1.nih.gov/grants/funding/women_min/guidelines_amended_10_2001.htm. Accessed 16 July 2013
7. Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. Lancet. 2002;359:966–70.
8. Williams OD. Fundamentals of clinical research (Power Point Lecture). 2005.
9. Practical aspects of decision making in clinical trials: the coronary drug project as a case study. The Coronary Drug Project Research Group. Control Clin Trials 1981;1:363–76.
10. DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. Stat Med. 1994;13:1341–52; discussion 53–6.
11. DeMets DL, Pocock SJ, Julian DG. The agonising negative trend in monitoring of clinical trials. Lancet. 1999;354:1983–8.
12. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer. 1976;34:585–612. PMC2025229.
13. Hulley S, Cummings S, Browner W. Designing clinical research. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 2000.
14. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol. 1971;44:793–7.
15. Pocock SJ. When to stop a clinical trial. BMJ. 1992;305:235–40.
16. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;35:549–56.
17. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. Am Stat. 1966;20:18–23. doi:10.1080/00031305.1966.10479786.
18. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton: Chapman and Hall; 2000.
19. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. J R Stat Soc Ser A Stat Soc. 1969;132:235–44.
20. McPherson K. The problem of examining accumulating data more than once. N Engl J Med. 1974;290:501–2.
21. Caiff RM, Ellenberg SS. Statistical approaches and policies for the operations of Data and Safety Monitoring Committees. Am Heart J. 2000;141:301–5.
22. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA. 2005;294:2203–9.
23. Monitoring Boards for Data and Safety. http://public.nhibi.nih.gov/ocr/home/GetPolicy.aspx?id=8. Accessed 14 June 2007.
24. Tharmananathan P, Calvert M, Hampton J, Freemantle N. The use of interim data and data monitoring committee recommendations in randomized controlled trial reports: frequency, implications and potential sources of bias. BMC Med Res Methodol. 2008;8:12.
25. Probstfield JL, Frye RL. Strategies for recruitment and retention of participants in clinical trials. JAMA. 2011;306:1798–9. doi:10.1001/jama.2011.1544.

Chapter 10
Meta-analysis, Evidence-Based Medicine, and Clinical Guidelines

Stephen P. Glasser and Sue Duval

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*
>
> R.A. Fisher. Presidential Address by Professor R.A. Fisher, Sc.D., F.R.S. Sankhyā: The Indian Journal of Statistics (1933–1960), Vol. 4, No. 1 (1938), pp. 14–17. http://www.jstor.org/stable/40383882.

Abstract Meta-analysis refers to methods for the systematic review of a set of individual studies (either from the aggregate data or the individual patient data) with the aim to quantitatively combine their results. This has become a popular approach to attempt to answer questions when the results from individual studies have not been definitive. This chapter will discuss meta-analyses and highlight issues that need critical assessment before the results of the meta-analysis are accepted. Some of these critical issues include: publication bias, sampling bias, and study heterogeneity. Evidence-based medicine and clinical practice guidelines are dependent upon meta-analyses to guide their recommendations. Evidence-based medicine is an apt term to the extent that it advocates more reliance on clinical research than on personal experience or intuition; and, has led to a paradigm outlining the "level of evidence" that addresses a particular clinical question (also see Chap. 3). These "levels of evidence" are also utilized by clinical practice guidelines, but "as the number of available guidelines provided by a variety of sources has literally exploded, serious questions and controversies have arisen about how guidelines should be developed, implemented, and evaluated."

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT 638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S. Duval, Ph.D.
Cardiovascular Division, University of Minnesota Medical School, Minneapolis, MN, USA
e-mail: sueduval@umn.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_10,

Keywords Meta-analysis • Publication bias • Systematic review • Evidence-based medicine • Clinical practice guidelines

Introduction

Meta- is from Latin meaning among, with, or after; occurring in succession to, situated behind or beyond, more comprehensive, or transcending. This has led some to question if meta-analysis is to analysis as metaphysics is to physics (metaphysics refers to the abstract or supernatural), to which a number of article titles would attest, such as: "is a meta-analysis science or religion?" [1]; "have meta-analyses become a tool or a weapon?" [2]; "meta-statistics: help or hinderance?" [3] and, "have you ever meta-analysis you didn't like?" [4], or as Bangalore put it "a meta-analysis is like a sausage. God and the butcher know what goes in it and neither would ever eat any" [5]. Overviews, systematic reviews, pooled analyses, quantitative reviews and quantitative analyses are other terms that have been used synonymously with meta-analysis, but some distinguish between them. For example, pooled analyses might not necessarily use the true meta-analytic statistical methods, and quantitative reviews might similarly be different than a meta-analysis. Compared to traditional reviews, meta-analyses are often more narrowly focused, usually examine one clinical question, and necessarily have a strong quantitative component. Meta-analysis can be literature based and these are essentially, studies of studies. Said simply, meta-analysis is the statistical combination of two or more separate studies, with the potential advantages being improved precision and increased power. The majority of meta-analyses rely on published reports, however more recently, meta-analyses of individual patient (participant) data (IPD) have appeared.

The earliest meta-analysis may have been that of Karl Pearson in 1904, which he applied in an attempt to overcome the problem of reduced statistical power in studies with small sample sizes [6]. The first meta-analysis of medical treatment is probably that of Henry K Beecher on the powerful effects of placebo, published in 1955 [7]. But, the term meta-analysis is credited to Gene Glass in 1976 [8]. Only four meta-analyses could be found before 1970, 13 were published in the 1970s and fewer than 100 in the 1980s. Since the 1980s more than 10,000 meta-analyses have been published. Why this popularity of meta-analysis, and why do a meta-analysis in the first place? Individual studies attempt to make inferences by setting up experimental contrasts that pertain to the hypothesis at hand. Nevertheless, observed findings are subject to random variation that could lead the inference astray, and it is also difficult to test the consistency of findings across a variety of settings from a single study. The goal of a meta-analysis is to enhance inference by increasing power and by assessing the consistency of findings across studies; and in so doing one can more appreciate the degree of uncertainty in the research question, and the degree of heterogeneity between studies.

Definition

Meta-analysis refers to methods for the systematic review of a set of individual studies or patients (subjects) within each study, with the aim to quantitatively combine their results. Meta-analysis has become popular for many reasons, some of which include**:**

– The adoption of evidence-based medicine which requires that all reliable information is considered
– The desire to avoid narrative reviews which are often misleading or inconclusive
– The desire to interpret the large number of studies that have been conducted about a specific intervention
– The desire to increase the statistical power of the results by combining many smaller sized studies

Some definitions of a meta-analysis include:

- An observational study in which the units of observation are individual trial results or the combined results of individual patients (subjects) aggregated from those trials.
- A scientific review of original studies in a specific area aimed at statistically combining the separate results into a single estimate
- A type of literature review that is quantitative
- A statistical analysis involving data from two or more trials of the same treatment and performed for the purpose of drawing a global conclusion concerning the safety and efficacy of that treatment

One should view the steps in designing a meta-analysis the same way as one views the steps take in designing a clinical trial (unless one is performing an exploratory meta-analysis), except that most meta-analyses are retrospective and observational. Beyond that, a meta-analysis is like a clinical trial except that the units of observation may be individual subjects within trials, or individual trial results. Thus, all the considerations given to the strengths and limitations of clinical trials should be applied to meta-analyses (e.g. a clearly stated hypothesis, a predefined protocol, considerations regarding selection bias, etc.).

The reasons one performs a meta-analysis is to 'force' one to review all pertinent evidence, to provide quantitative summaries, to integrate results across studies, and to provide for an overall interpretation of these studies. This allows for a more rigorous review of the literature, and it increases sample size and thereby potentially enhances statistical power. That is to say, the primary aim of a meta-analysis is to provide a more precise estimate of an outcome (say a medical therapy in reducing mortality or morbidity) based upon a weighted average of the results from the studies included in the meta-analysis (Table 10.1). The concept of a 'weighted average' is an important one. In the most basic approach, the weight given to each study is the inverse of the variance of the effect; that is, on average, the smaller the variance, and the larger the study, the greater the weight one places on the results of that study. Because the results from different studies investigating different but hopefully similar questions are often measured on different scales, the dependent variable in a

Table 10.1 Some reasons to perform a meta-analysis

"Force" a rigorous literature review
Resolve uncertainty when reports disagree
Increase sample size
Enhance statistical significance of subgroup analyses
Enhance scientific credibility of some observations
May identify new research directions
May help put into focus a controversial study
Provide more precise effect size estimates
Allow one to assess variability between studies
Increase statistical power
May identify characteristics associated with particularly effective treatments
Allow for study of heterogeneity

Table 10.2 Comparison of expert reviews vs. meta-analysis

	Expert review	Meta-analysis
Question	Broad	Focused
Sources	Often not specified	Comprehensive
Search	Ad-hoc	Explicit
Selection	Often not specified	Criterion-based
Appraisal	Variable	Rigorous
Synthesis	Usually qualitative	Qualitative or quantitative
Inference	Sometimes evidence-based	Usually evidence-based

meta-analysis is typically some standardized measure of effect size. In addition, meta-analyses may enhance the statistical significance of subgroup analysis, and enhance the scientific credibility of certain observations.

Finally, meta-analyses may identify new research directions or help put into focus the results of a controversial study. As such, meta-analyses may resolve uncertainty when reports disagree, improve estimates of effect size, and answer questions that were not posed at the start of individual trials, but are now suggested by the trial results. Thus, when the results from several studies disagree with regard to the magnitude or direction of effect, or when sample sizes of individual studies are too small to detect an effect, or when a large trial is too costly and/or too time consuming to perform, a meta-analysis should be considered.

One should make the distinction between meta-analysis, a systematic review, and an expert review. Meta-analysis is quantitative and employs statistical methods to combine and summarize the results of several studies; a systematic review is the process for searching the literature appropriately, in order to find the relevant information. Expert reviews are broad and frequently biased summaries by a leading authority in a given field (Table 10.2).

Weaknesses

As is true for any analytical technique, meta-analyses have weaknesses. For example, they are sometimes viewed as more authoritative than is justified. After all, meta-analyses are retrospective repeat analyses of prior published data. Rather, meta-analyses should be viewed as nearly equivalent (if performed properly under rigid study design characteristics) to a large, multi-center study. In fact, meta-analyses are really studies in which the 'observations' are not under the control of the meta-investigator (because they have already been performed by the investigators of the original studies); the included studies have not been obtained through a randomized and blinded technique; and, one must assume that the original studies have certain statistical properties they may not, in fact, have. In addition, one must rely only on reported rather than directly observed values, unless an IPD meta-analysis is undertaken.

There are at least nine important considerations in performing or reading a meta-analysis (Table 10.3):

1. They are sometimes performed to confirm an observed trend (this is equivalent to testing before hypothesis generation)
2. The sample of studies included in a meta-analysis may not be representative
3. Publication bias
4. Difficulty in pooling across different study designs
5. Dissimilarities of control treatment
6. Differences in the outcome variables
7. Studies are reported in different formats with different information available
8. The issues surrounding the choice of fixed versus random modeling
9. Alternative modeling

Table 10.3 At least nine considerations when performing or reading a meta-analysis

Is it being done to confirm observed trends?
Pooling across studies is difficult
Sample bias
Publication bias
Control treatment dissimilarities
Differences in primary and secondary outcomes across studies
Differences in reporting outcomes
Weighting
Modeling

Meta-analyses are Sometimes Performed to Confirm Observed Trends (i.e. Testing Before Hypothesis Generation)

Frequently in meta-analyses, the conduct of the analysis is to confirm observed 'trends' in sets of studies; and, this is equivalent to examining data to select which statistical analyses should be performed, rather than the reverse. This is well known to introduce spurious findings. It is important to be hypothesis driven i.e. to perform planning steps in the correct order (if possible).

In planning the meta-analysis, the same principles apply as planning any other study. That is, one forms a hypothesis, defines inclusion and exclusion criteria, collects data, tests the hypothesis, and reports the results. But, as previously mentioned, just like other hypothesis testing, the key is to avoid spurious findings by keeping these steps in the correct order, and this is sometimes NOT the case for meta-analyses. For example, frequently the 'trend' in the data is already known; in fact, most meta-analyses are performed because of a suggestive trend. In Petitti's steps in planning a meta-analysis she suggests first addressing the objectives (i.e. state the main objectives, specify secondary objectives); perform a review; information retrieval; specify MEDLINE search criteria; and explain approaches to capture 'fugitive' reports (those not listed in MEDLINE or other search engines and therefore not readily available) [9].

The Sample of Studies Included in a Meta-analysis May Not Be Representative

As with sampling in clinical trials identifying studies to be considered for inclusion is in essence, defining the 'sampling frame' for the meta-analysis. The overall goal is to include all pertinent studies; and, several approaches are possible. One approach could be: 'I am familiar with the literature and will include the important studies'. With this approach, there may be a tendency to be aware of only certain types of studies and selection will therefore be biased. A more scientific and valid approach is where one uses well-defined criteria for inclusion and exclusion applying an objective screening (search) tool such as MEDLINE. Clearly defined keywords and MESH terms, clearly defined years of interest, and a transparent description of what the meta-investigator did must be included in any report. Also, the impact of the 'Search Engine' on identifying papers must be adequately reported to allow for study replication. Surprising to some is that there may be problems with using MEDLINE alone to screen for articles. Other searches can be done with EMBASE or PUBMED and seeking the help of a trained Biomedical Librarian is generally advisable. In addition, not all journals are included in these search engines and there is dependence on keywords assigned by authors and MESH terms by Medline indexers [10]. Further, searches may not include fugitive or grey literature, government reports, book chapters, proceedings of conferences, published dissertations, etc.

As previously stated, the included studies in a meta-analysis have not been obtained through a randomized and blinded technique, so that selection bias becomes an issue. Selection bias occurs because studies are 'preferentially' included and excluded and these decisions are influenced by the meta-investigators prior beliefs as well as the fact that studies are included based upon recognized 'authorities'. That is, investigator bias occurs because the investigators who conducted the individual studies included in the meta-analysis may have introduced their own bias.

It is necessary for a complete meta-analysis to go to supplemental sources for studies, such as studies of which authors are personally aware, studies referenced in articles retrieved by Search Engines, and searches of Dissertation Abstracts to name a few. The biggest limitation, however, is how to search for unpublished and unreported studies. This latter issue is clearly the most challenging (impossible?), and opens the possibility for publication bias and the "file-drawer" problem.

Publication Bias (and the File-Drawer Problem)

Publication bias is one of the major limitations of meta-analysis as it derives from the fact that for the most part, studies that are published have positive results, so that negative studies are underrepresented and if published take longer to appear in the literature ("pipeline effect"). Stated another way, publication bias results from the selective publication of studies based on the direction and magnitude of their results. As an example, Turner et al. found that 17 % of 24 FDA registered trials were unpublished and 3 of 4 of the unpublished trials failed to show benefit over placebo [11].

The pooling of results of published studies alone can lead to an overestimation of the effectiveness of the intervention, and the magnitude of this bias tends to be greater for observational studies compared to RCTs. In fact, positive studies are three times more likely to be published than negative ones and this ratio is even greater for observational studies. Thus, investigators tend not to submit negative studies (this is frequently referred to as the 'file-drawer' problem), journals do not publish negative studies as readily, funding sources may discourage publication of negative studies, negative studies that do get published are published in lower impact journals some of which might not be indexed in Medline or other databases. One also has to be wary of overrepresentation of positive studies because duplicate publication can occur. The scenario resulting in publication bias goes something like this: one thinks of an exciting hypothesis, examines the possibility in existing data, if significant, the findings are published, but if non-significant the investigator loses interest and buries the results (i.e. puts them in a file drawer). Even if one is 'honorable' and attempts to publish a non-significant study, often the editor/reviewer will bury the result for you, since negative results are difficult to publish. One then continues on to the next idea and forgets that the analysis was ever performed. The obvious result of this is that the literature is more likely to include mostly positive findings and thereby is biased toward benefit. Publication bias is equivalent to performing a screen to select patients who only respond positively to a treatment before performing a clinical trial to examine the efficacy of that treatment.

To moderate the impact of publication bias, one attempts to obtain all published and unpublished data on the question at hand. There are also tests for the presence of publication bias, and methods to estimate the impact of publication bias and adjust for it. It should be noted that publication bias is a greater problem in epidemiological studies than clinical trials, since it is difficult to perform a major RCT and not publish the results even if negative, while for epidemiologic studies negative results are much less likely to be published.

As mentioned, there are ways that one can determine the likelihood that publication bias is influencing the meta-analysis. One of the simplest methods is to construct a funnel plot, which is a scatter plot of individual study effects against a measure of precision within each study. In the absence of bias, the funnel plot should depict an inverted 'funnel' shape centered about the true overall mean which the meta-analysis is trying to estimate. This is because we expect a wider spread of effects among the smaller studies. If the funnel appears truncated, it is likely that a group of studies is missing from the analysis set. It should be kept in mind however that publication bias is but one potential reason for this 'funnel plot asymmetry', and for this reason, current practice is to consider other mechanisms for the missing studies, such as English language bias, clinical heterogeneity, and location bias to name a few [12].

There are a number of relatively simple quantitative methods for detecting publication bias in the literature, including the rank correlation test of Begg and the regression-based test of Egger et al. [13, 14]. The Trim and Fill method can be used to estimate the number of missing studies and to provide an estimate of the treatment effect after adjustment for this bias [15]. The mechanics of this approach are displayed in Fig. 10.1a, using a meta-analysis of the effect of gangliosides and mortality from acute ischemic stroke [16]. Although in this example, the effect size is not great, the striking aspect of the plot is that it appears that there are no negative effects of therapy. The question is whether that observation is true or if this is an example of publication bias where the negative studies are not represented. Figure 10.1b shows what happens when the asymmetric studies are 'trimmed' to generate a symmetric plot to allow estimation of the true pooled effect (in this example, the five rightmost studies are trimmed). These trimmed studies are then returned, along with their imputed or 'filled' symmetric counterparts. An adjusted pooled estimate and corresponding confidence interval are then calculated based on the now presumed complete dataset (bottom panel). The authors of this method stress that the main goal of such an analysis is to allow a 'what if' approach; that is, to allow sensitivity analyses to the missing studies, rather than actually finding the values of those studies per se. Heterogeneity, reporting bias, and chance may all lead to asymmetry or other shapes in funnel plots (box). Funnel plot asymmetry may also be an artifact of the choice of statistics being plotted. Reporting biases arise when the dissemination of research findings is influenced by the nature and direction of results. As noted by Sterne et al. [12], positive studies are more likely to be published, published

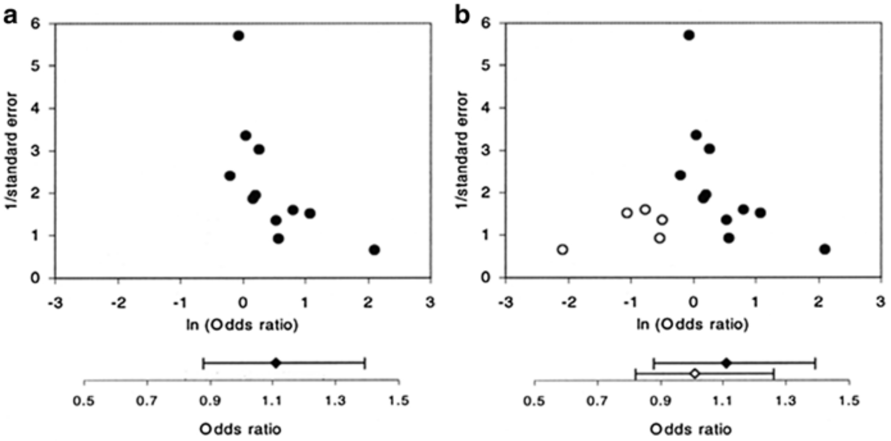


Fig. 10.1 (**a**) A of the studies included in the meta-analysis. (**b**) Filled "presumed" negative studies shown as unfilled circles, with the adjusted odds ratio calculated

rapidly, published in English, published more than once, published in high impact journals, and cited by others; while negative studies may be filtered, manipulated, or presented in such a way that they become positive. Reporting biases can have three types of consequence for a meta-analysis:

- A systematic review may fail to locate an eligible study because all information about it is suppressed or hard to find (publication bias)
- A located study may not provide usable data for the outcome of interest because the study authors did not consider the result sufficiently interesting (selective outcome reporting)
- A located study may provide biased results for some outcome—for example, by presenting the result with the smallest P value or largest effect estimate after trying several analysis methods (selective analysis reporting).

These biases may cause funnel plot asymmetry if statistically significant results suggesting a beneficial effect are more likely to be published than non-significant results. Such asymmetry may be exaggerated if there is a further tendency for smaller studies to be more prone to selective suppression of results than larger studies. This is often assumed to be the case for randomized trials. For instance, it is probably more difficult to make a large study disappear without a trace, while a small study can easily be lost in a file drawer. The same may apply to specific outcomes. For example, it is difficult not to report on mortality or myocardial infarction if these are outcomes of a large study. Smaller studies have more sampling error in their effect estimates. Thus even though the risk of a false positive significant finding is the same, multiple analyses are more likely to yield a large effect estimate that may seem worth publishing. However, biases may not act this way in real life; funnel plots could be symmetrical even in the presence of publication bias or selective outcome reporting for example, if the published findings point to effects in different

Table 10.4 Publication of studies based upon positive vs. negative results

Publication status of studies reviewed by the Central Oxford Research Ethics Committee (1984–1987)

	Statistically significant	Statistically non-significant
% Published	60	35
% Only presented	45	22
% Neither	15	42

Adapted from Easterbrook [17]

Table 10.5 Magnitude of effect size in published vs. registered studies

Meta-analysis of published vs. registered studies of treatment with alkylating agents for advanced ovarian cancer

	Published studies (n = 16)	Registered studies (n = 13)
Survival ratio	1.16	1.06
95 % CI	1.06–1.27	0.97–1.15
P-Value	0.02	0.24

Adapted from Easterbrook [17]

directions but unreported results indicate neither direction. Alternatively, bias may have affected few studies and therefore not cause glaring asymmetry.

Perhaps the best approach to avoid publication bias is to have a registry of all trials at their inception, that is, before results are available, thereby eliminating the possibility that the study results would influence inclusion into the meta-analysis. After a period of apathy, this concept is taking hold and a website (*clinicaltrials.gov*) is now available. But, to emphasize the importance of this, Table 10.4 points out an example of the publication status of studies that were statistically significant vs. those that were not; and Table 10.5 emphasizes the magnitude of outcome bias seen in this set of published vs. registered studies.

The effect of publication bias on meta-analytical outcomes was demonstrated by Glass et al. in 1979 [18]. They reported on 12 meta-analyses, and in every instance where it could be determined, found that the average experimental effect from studies published in journals was larger than the corresponding effect estimated from unpublished work (mostly from theses and dissertations), accounting for almost a 33 % bias in favour of the benefit. As a result, some have suggested that a complete meta-analysis should include attempts to contact experts in the field as well as authors of referenced articles for access to unpublished data. More recent estimates have suggested that the effect of publication bias accounts for 5–15 % in favour of benefit.

Some literature that is available but hard to find includes grey and fugitive literature. Grey literature refers to a body of materials that cannot be found easily through conventional channels, "but which is frequently original and usually recent". The "Grey Information Functional Plan," defines grey literature as foreign or domestic open source material that usually is available through specialized channels and may not enter normal channels or systems of publication, distribution, bibliographic control, or acquisition by booksellers or subscription agents. Examples of grey literature include technical reports from government agencies or scientific research groups, working papers from research groups or committees, and white papers. But, the identification and acquisition of grey literature poses difficulties for librarians and other information professionals for several reasons. Generally, grey literature lacks strict bibliographic control, meaning that basic information such as author, publication date or publishing body may not be easily discerned. Similarly, non-professional layouts and formats and low print runs of grey literature make the organized collection of such publications challenging compared to more traditional published media such as journals and books. Fugitive literature is literally the ones for which you have to hunt. On the World Wide Web, it is not always easy to hunt for specific information, particularly if you do not know where to begin. The following provides a partial list of websites that provide entry points for searching fugitive literature:

- http://www.google.com Meta search engine that searches across other engines
- http://www.healthfinder.gov Healthcare information from the USDHHS
- http://www.guidelines.gov/index.asp Summary Guidelines info from the AHRQ
- http://www.cdc.gov Healthcare information from the CDC

The Difficulty in Pooling Across a Set of Individual Studies and Heterogeneity

One of the reasons that it is difficult to pool studies is selection bias. Selection bias occurs because studies are 'preferentially' included and excluded and these are influenced by the meta-investigators prior beliefs as well as the fact that studies included are based upon recognized 'authorities'. That is, this type of bias occurs because the investigators who conducted the individual studies included in the meta-analysis may have introduced their own bias. In addition, there is always a certain level of heterogeneity of study characteristics included in a given meta-analysis so that as the cliché goes 'by mixing apples and oranges with an occasional lemon, ones ends up with an artificial product.' Glass argued this point rather eloquently as follows:

> '*…Of course it mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavor worthy of true scientists; comparing apples to apples is trivial.…*'
>
> *The same persons arguing that no two studies should be compared unless they were studies of the 'same thing' are blithely comparing persons within studies i.e. no two things can be compared unless they are the same…but if they are the same then they are not two things.'* Glass went on to use the classic paradox of Theseus's ship, which set sail on a 5-year journey. After nearly 5 years, every plank had been replaced. The question then i*s 'are Theseus and his men still sailing the ship that was launched 5 years earlier? What if as each plank was removed, it was taken ashore and repositioned exactly as it had been on the waters so that at the end of 5 years, there exists a ship on shore, every plank of which once stood exactly as it had been 5 years before. Is this new ship Theseus's ship, or is it the one still sailing?' The answer depends on what we understand the concept of 'same' to mean.*

Table 10.6 Some causes of heterogeneity

Differences in inclusion/exclusion criteria of the individual studies
Different control or treatment interventions (dose, timing, brand), outcome measures and definition, and different follow-up times
The reasons for withdrawals, drop-outs, cross-overs will likely differ between individual studies, as will the baseline status of the patients and the settings
The quality of the study design and its execution will likely differ

> *Glass goes on to consider the problem of the persistence of personal identity when he asks the question 'how do I know that I am the same person who I was yesterday, or last year…?'*
>
> Glass notes that probably there are no cells that are in common between the current organism called Gene Glass and the organism 40 years ago by the same name [19].

Recall that a number of possible outcomes and interpretations of clinical trials is possible. When one trial is performed, the outcome may be significant, and one concludes that a treatment is beneficial, or the results may be inconclusive leading one to say that there is not convincing statistical evidence to support a treatment benefit. But when multiple trials are performed other considerations present themselves. For example, when 'most' studies are significant and in the same direction one can conclude a treatment is beneficial, but when 'most' studies are significant in different directions one might question whether there are differences in the population studied or methods performed that warrant further consideration. The question that may then be raised is 'Could we learn anything by combining the studies?' It is this latter question that is the underlying basis for meta-analysis. Thus, when there is some treatment or exposure under consideration we assume that there is a 'true' treatment effect that is shared by all studies, and that the average has lower variance than the data themselves. We then consider each of the individual studies as one data point in a 'mega-study' and presume that the best (most precise) estimate of this 'true' treatment effect is provided by 'averaging' across studies. But, when is it even reasonable to combine studies? The answer to this latter question is that studies must share characteristics, including similar 'experimental' treatment or exposure, similar 'standard' treatment or lack of exposure, similar follow-up protocol, outcome(s) and patient populations. It is difficult to pool across different studies, even when there is an apparent similarity of treatments. This leads to heterogeneity when one performs any meta-analysis. The causes of study heterogeneity are numerous. Some of them are (Table 10.6):

– Differences in inclusion/exclusion criteria of the individual studies comprising the meta-analysis
– Different control or treatment interventions [dose, timing, brand], outcome measures and definition, and different follow-up times were likely to be present in each individual study

- The reasons for withdrawals, drop-outs, cross-overs will likely differ between individual studies, as will the baseline status of the patients and the settings for each study.
- Finally, the quality of the study design and its execution will likely differ

Heterogeneity of the studies included in the meta-analysis can be tested. For example, Cochran's Q is a test of homogeneity that evaluates the extent to which differences among the results of individual studies are greater than one would expect if all studies were measuring the same underlying effect and the observed differences between them were due only to chance. A measure of the proportion of variation in individual study estimates that is due to heterogeneity rather than sampling error, (known as I^2), is available and is the preferred method of describing heterogeneity [20]. This index does not depend on the number of studies, the type of outcome data or the choice of treatment effect. I^2 is related to Cochran's Q statistic and lies between 0 and 100 %, making it useful for comparison across meta-analyses. Most reviewers consider that an I^2 greater than 50 % indicates heterogeneity between the component studies. Rather sensitivity analysis to differences in study quality is more common. Sensitivity analysis describes the robustness of the results by excluding some studies such as those for example, of greater risk of bias and/or smaller studies.

Dissimilarities in Control Groups

Just as important as the similarity in treatment groups, is that one needs to take great caution to ensure that control groups between studies included in the meta-analysis are similar. For example, one study in a meta-analysis may have a statin drug vs. placebo, while another study compares a statin drug plus active risk factor management (smoking cessation, hypertension control, etc.) compared to placebo plus active risk factor management. Certainly, one could argue that the between study control groups are not similar (clearly they are not identical), and one can only surmise the degree of bias that would be introduced by including both in the meta-analysis.

Heterogeneity in Outcome

One might expect that the choice of an outcome to be evaluated in a meta-analysis is a simple choice. In many meta-analyses, it is not as simple as one would think. For example, consider a meta-analysis shown in Table 10.7. The range of effect has a risk differential from an approximately 60 % decrease to 127 % increase. One should reasonably ask whether the studies included in the meta-analysis should demonstrate approximately consistent results. Does it make sense to combine studies that are significant in different directions? If studies provide remarkably different estimates of treatment effect, what does an average mean? This particular scenario is used to further illustrate the use of sensitivity analyses in meta-analysis.

Table 10.7 Meta-analysis of stroke as a result of an intervention

Study	Estimate (95 % CI)	
1	1.12 (0.79–1.57)	Fatal and nonfatal first stroke
2	1.19 (0.67–2.13)	Hospitalized F/NF stroke
3	1.16 (0.75–1.77)	Occlusive stroke
4	0.64 (0.06–6.52)	Fatal SAH
5	2.27 (1.22–4.23)	Fatal and nonfatal stroke or TIA
6	0.40 (0.01–3.07)	Fatal stroke
7	0.97 (0.50–1.90)	Fatal and nonfatal first stroke
8	0.63 (0.40–0.97)	Fatal occlusive disease
9	0.97 (0.65–1.45)	Fatal and nonfatal stroke
10	0.65 (0.45–0.95)	Fatal and nonfatal first stroke
OVERALL	0.96 (0.82–1.13)	

A so-called ‘influence analysis’ is derived in which the meta-analysis is re-estimated after omitting each study in turn. It may be reasonable to consider excluding particular studies, or to present the results with one or two studies included and then excluded. Many analyses start out with the intention of producing quantitative syntheses, and fall short of this goal [21]. If the reasons are well argued, this can often be the most reasonable outcome.

Studies are Reported in Different Formats with Different Information Available

Since studies are reported in different formats with different information available, the abstraction of data can become problematic. There is no reason to anticipate that investigators will report data in a consistent manner. Frequently, differences in measures of association (odds ratio versus regression coefficients versus risk ratios, etc.) are presented in different reports which then forces the abstractor to try to reconstruct the same measure of association across studies. When abstracting information for meta-analyses, one must go through each study and attempt to collect the information in the same format. That is, one needs either a measure of association (e.g. an odds ratio) with some measure of dispersion (e.g. variance, standard deviation, confidence interval), or cell frequencies in 2×2 tables. If one wants to present a meta-analysis of subgroup outcomes, pooling may be even more problematic than pooling primary outcomes. This is because subgroups of interest are frequently not presented in identical categories.

The issue of consistency in the reporting of studies is a particular problem for epidemiological studies where confounders are a major issue. Although confounders are easily addressed by multivariable models, there is no reason to assume that authors will use the same models in adjusting for confounders. Another related problem is the possibility that there are multiple publications from a single population, and it is not always clear that this has occurred. For example, let's say that there is a publication reporting results in 109 patients. Three years later a report from the same or similar authors reports the results of a similar intervention in 500 patients. The question is, were the 500 patients all new, or did the first report of 109 patients get included in the 500 now being reported?

The Use of Random vs. Fixed Analysis Approaches

By far, the most common approach to weighting the results in meta-analyses is to calculate a 'weighted average' of the effects (e.g. odds ratios, risk ratios) across the studies. This has the overall goal of:

- Calculating an 'weighted average' measure of effect, and
- Performing a test to see if this estimated effect is different from the null hypothesis of no effect

In considering whether to use the fixed effect or random effects modeling approach, the fixed effect approach assumes that studies included in the meta-analysis are the only studies to which the inference will be applied, while the random effects approach assumes that the studies are a random sample of studies that may have occurred, and inference can be extended to "studies like these". The fixed effect model weights the studies by their 'precision' only. Precision is largely driven by the sample size and reflected by the widths of the 95 % confidence limits about the study-specific estimates. In general, when weights are assigned by the precision of the estimates they are proportional to (1/var(study)). This method assigns a bigger weight to a big and poorly-done study than it does to a small and well-done study. Thus, a meta-analysis that includes one or two large studies is largely a report of just those studies. Random effects models estimate a between study variance component, and incorporate that into the model. This effectively makes the contributions of individual studies to the overall estimate more uniform. It also increases the width of the confidence interval of the overall effect. The random effects approach is likely more representative of the underlying statistical framework and the use of the 'fixed' approach can provide an underestimate of the true variance and may falsely inflate power to see effects. Most older meta-analyses have used the fixed effect approach, while many newer meta-analyses are using the random effects approach since it is more representative of the 'real' world. A reasonable approach is to present the results from both models.

Assignment of Weights

Alternative weighting schemes have been suggested, such as weighting by the quality of the study, with points given based on the number of variables [22]. The problem with weighting is that one has started the meta-analysis in order to have an objective method to combine studies to provide an overall summary, and with weighting we are subjectively assigning weights to factors so that we can objectively calculate a summary measure. However, this aforementioned weighting is but one scheme and its use has been questioned by many experts in the field. Most meta-investigators now use fixed, random, or Bayesian approaches [23].

Statistical and Graphical Approaches

Forest Plot

The forest plot is a common graphical way of portraying the data in a meta-analysis. In this plot, the point is the estimate of the effect, the size of the point is proportional to the size of the study, and the confidence intervals around that point estimate are displayed (for example, an odds ratio of 1 means the outcome is not affected by the intervention under study). In Fig. 10.2, a hypothetical forest plot of log hazard ratios for each study, ordered by the size of the effect within each study is shown. At the bottom, a diamond shows the combined estimate from the meta-analysis.

An example of some of these aforementioned principles is demonstrated in a theoretical meta-analysis of six studies. For this ‘artificial’ meta-analysis, only multicenter randomized trials were included, and the outcome is total mortality. Tables 10.8a, 10.8b, and 10.8c, present the raw data, mortality rates and odds ratios.

The fundamental statistical approach in meta-analysis is similar to that of an RCT in that the hypothesis is conceived to uphold the null. According to the Mantel-Haenszel-Peto method, a technique commonly used when events are sparse, a 2×2 table is constructed for each study to be included, and the observed number for the outcome of interest is computed [24]. From that computation one subtracts the expected outcome had no intervention been given. If the intervention of interest has no effect, the observed minus the expected should be about zero; if the intervention is favorable (with the measure of association being the odds ratio-OR) the OR will be greater than 1 (as will its confidence limits). The magnitude of effect can be calculated in meta-analyses using a number of measures of association, such as the odds ratio (OR), relative risk (RR), risk difference (RD), and/or the hazard ratio (HR), to name a few. The choice is, to a great degree, subjective as discussed in Chap. 16, and briefly in section “Studies are reported in different formats with different information available” above.

One limited type of meta-analysis, and a way to overcome some of the limitations of meta-analysis in general, is to preplan them with the prospective registration of studies, as has been done with some drug developments. Berlin and Colditz present the potential uses of meta-analyses (primarily of RCTs) in the approval and postmarketing evaluation of approved drugs [25]. If a sponsor of a new drug has a program to conduct a number of clinical trials, and the trials are planned as a series with prospective registration of studies at their inception, one has a focused question (e.g. drug efficacy for lowering the total cholesterol), all patients are included (so no

Table 10.8a The raw data from the six studies included in the meta-analysis

Raw data						
	Treatment A			PLACEBO		
Study	Total no. of patients	No. dead	No. alive	Total no. of patients	No. dead	No. alive
1	615	49	566	624	67	557
2	758	44	714	771	64	707
3	317	27	290	309	32	277
4	832	102	730	850	126	724
5	810	85	725	406	52	354
6	2267	246	2021	2257	219	2038
Total	5599	553	5046	5217	560	4657

Table 10.8b The individual mortality rates from the six studies included in the meta-analysis

	ASPIRIN	PLACEBO	Aspirin-Placebo		
Study	Mortality rate	Mortality rate	Diff	SE of diff	P-value
1	.0797	.1074	−.0277	.0165	0.047
2	.0580	.0830	−.0250	.0131	0.028
3	.0852	.1036	−.0184	.0234	0.216
4	.1226	.1482	−.0256	.0167	0.062
5	.1049	.1281	−.0231	.0198	0.129
6	.1085	.0970	.0115	.0090	0.898

Table 10.8c The odds ratios from the six studies included in the meta-analysis

Odds ratios for the six trials				
Study	Log odds ratio	SE [log OR]	Odds ratio	CI on OR
1	−0.33	0.197	0.72	[0.49,1.06]
2	−0.38	0.203	0.68	[0.46,1.02]
3	−0.22	0.275	0.81	[0.47,1.38]
4	−0.22	0.143	0.80	[0.61,1.06]
5	−0.23	0.188	0.80	[0.55,1.15]
6	0.12	0.098	1.13	[0.93,1.37]

publication bias occurs), one then has the elements of a well-planned meta-analysis. In Table 10.9, Berlin and Colditz present their comparison of trials as they relate to four key elements of several types of clinical trials [23].

In designing a meta-analysis (or reading one in the literature) one should be certain that a number of details are included so the validity of the results can be weighed. Some of the considerations are: listing the trials included and excluded in the meta-analysis and the reasons for doing so; clearly defining the treatment assignment in each of the trials; describing the ranges of patient characteristics, diagnoses, and treatment assignment; and, addressing what criteria were used to decide that the studies analyzed were similar enough to be pooled. Finally, meta-analyses can provide more precise estimates of the effects of interventions, increase

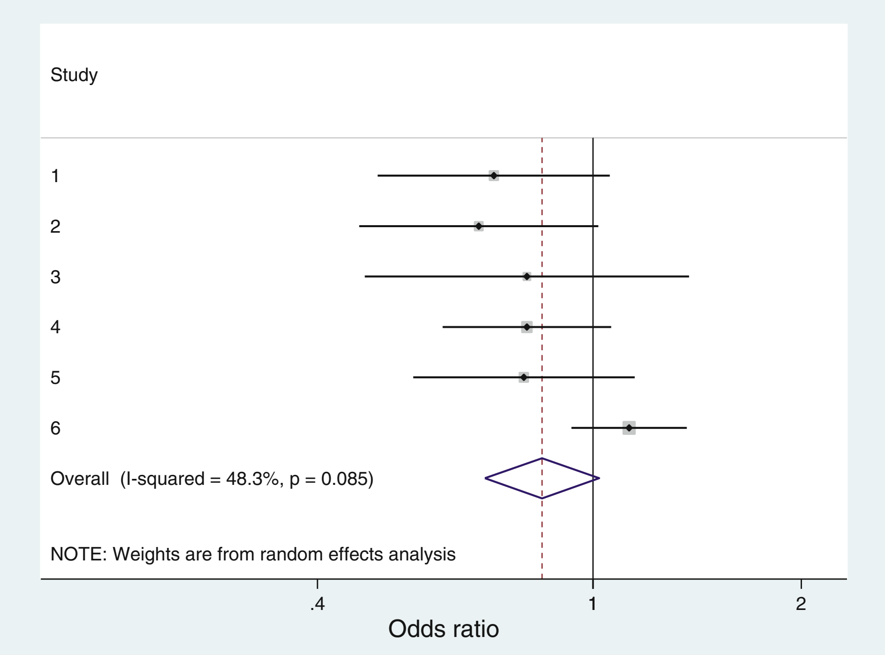


Fig. 10.2 Example of a forest plot

Table 10.9 Variables relating to publication bias, generalizability, and validity with different study approaches

Approach	Avoids publication bias	Generalizes across protocols	Generalizes across centers	Validity
Pre-planned	+++	+++	+++	++
LST	++	–	+++	++
Retrospective	–	++	++	+
2 RCTs	–	++	++	++
1 RCT	–	–	–	+

LST Large Simple Trial

statistical power, assess the amount of variability between studies, reach agreements when results from different studies are discordant, and identify study characteristics associated with particularly effective treatments (Table 10.10). Typically, analyses should include: the point estimate, 95 % confidence limits, a graphical display (forest plot), p values, a statistical test for heterogeneity, sensitivity analyses, and potential sources of bias (e.g. publication bias using the funnel plot).

As is true for clinical trials and the CONSORT Guidelines, there are guidelines for the reporting of meta-analyses: PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) is an update of QUORUM (QUality Of Reporting of Meta-Analyses), for meta-analyses of RCTs; and, Meta-analysis Of Observational

Table 10.10 What can meta-analyses provide?

Provide more precise estimates of the effects of interventions
Increase statistical power
Assess the amount of variability between studies
Reach agreements when results from different studies are discordant
Identify study characteristics associated with particularly effective treatments

Studies in Epidemiology (MOOSE) for meta-analyses of Observational studies see Chap. 2 [26]. In addition, a critical appraisal checklist for a systematic review has been developed under the guidance of the Critical Appraisal Skills Program [27].

Evidence-Based Medicine

> *'It ain't so much what we don't know that gets us into trouble as what we do know that ain't so' (Will Rogers)* (http://humrep.oxfordjournals.org)

Clinical Effectiveness, Clinical Governance, Risk Management, Benchmarking—Essence of Care, NHS Knowledge and Skills Framework, and Evidence-based Practice are but a few of the terms that have now become part of everyday practice for health professionals. Such terms appear to be open to interpretation and confusion. Evidence-based medicine was originally defined as the process of *"…integrating individual clinical expertise and the best available external clinical evidence from systematic research."* [28] Meta-analysis and evidence-based medicine (EBM) arose together as a result of the fact that the traditional way of learning (the Historic Paradigm i.e. 'evidence' is determined by the leading authorities in the field from textbooks, review articles, seminars, and consensus conferences) was based upon the assumption that experts represented infallible and comprehensive knowledge. Numerous examples of the fallibility of that paradigm are present in the literature e.g.:

- Prenatal steroids for mothers to minimize risk of Respiratory Distress Syndrome (RDS)
- Treatment of eclampsia with magnesium sulfate vs. diazepam
- NTG use in suspected MI
- The use of diuretics for pre-eclampsia

In 1979 Cochrane stated 'It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or sub-specialty, updated periodically, of all relevant randomized controlled trials' [29]. The idea of EBM then was to devise answerable questions, track down the best evidence to answer them, critically appraise the validity and usefulness of the evidence, apply the appraisal to clinical practice, and to evaluate one's performance after applying the evidence into practice [30]. As such, EBM called for the integration of individual clinical expertise with the best available external evidence from systematic research (i.e. meta-analysis). One definition of EBM is the conscientious, explicit judicious use of current best available evidence in making decisions about the care of individual patients with the use of RCTs, wherever possible, as the gold standard [31]. EBM also incorporates the need to encourage patterns of care that do more good than harm.

It has been said, it is not that we are reluctant to use evidence-based approaches, it is that we may not agree on what the evidence is, so why shift to an EBM approach? The answers are many, but include the fact that the volume of new evidence can be overwhelming (this remains the clinician's biggest challenge), there is limited time available to keep up, up-to-date knowledge and clinical performance deteriorate with time; and, traditional CME has not been shown to improve clinical performance.

The necessary skills for EBM include the ability to precisely define a patient problem, ascertain what information is required to resolve the problem, the ability to conduct an efficient search of the literature with the selection of the most relevant articles, the ability to determine a study's validity, extract the clinical message and apply it to the patient's problem [32].

There are, of course criticisms of the EBM approach. For example, some feel that evidence is never enough i.e. evidence alone can never guide our clinical actions and that there is a shortage of coherent, consistent scientific evidence. Also, the unique biological attributes of the individual patient render the use of EBM to that individual, at best, limited. For many, the use of EBM requires that new skills be developed in an era of limited clinician time and technical resources. Finally, who is to say what the evidence is or that evidence-based medicine works? Some have asked, are those who do not practice EBM practicing 'non-evidence-based medicine'? Karl Popper perhaps summarized this best in a very thoughtful and insightful commentary, where he discussed the differences between evidence, truth, and knowledge when he noted that there are all kinds of sources of our knowledge but none has authority [33, 34]. *"Evidence is information that is used to approach truth, whereas truth is an infallible, unequivocal, immutable fact. The definition of knowledge…is typically used as a representation of a person's comprehension of a particular subject."* He further notes that *although truth is our ultimate desire it is likely unattainable, and that although evidence imbues us with knowledge, it does not affirm truth."* [34] As an example, RCTs use inductive reasoning to draw conclusions that are expressions of probability (not truth), but are often dubbed as truth. Baum cites Prasad et al. who define to "*signify the phenomenon of a new trial-superior to predecessors because of better design, increased power, or more appropriate controls-contradicting current clinical practice*." [35] In Baum's study 212 original publications in the New England Journal of Medicine were reviewed, 124 of which made some claim with respect to medical practice. Of these 124 there were 16 reversals (13 %). That is "truth" was reversed 13 % of the time [35].

Baum ends with the following "…*through the medical systems endowment of the P value and the RCT with boundless unfounded power, the lay public and physicians alike have become confused. Conflicting publications are released nearly on a weekly basis, each of them being treated as gospel with its message being shouted from the rooftops by the media as well as the camera-adoring members of our profession.*

The fact that science is a process is ignored. Undecipherable statistical jargon cloaks the fact that medical evidence emanates not from truth, but instead the falsifiable proof (the rejection of the null hypothesis)." [35]

Evidence-Based Medicine is perhaps a good term to the extent that it advocates more reliance on clinical research than on personal experience or intuition. But, medicine has always been taught and practiced based upon available scientific interpretation. The question can then be asked is whether the results of a clinical trial hardly deserve the title *evidence* as questions arise about the statistical and design aspects, and data analysis, presentation, and interpretation contain many subjective elements as we have discussed in prior chapters. Thus, even if we observe consistency in the results and interpretation (a rare occurrence in science) how many times should a successful trial be replicated to claim proof? That is, whose evidence is *the evidence in evidence-based medicine*?

The five steps of EBM were first described in 1992 as follows [36]

1. The translation of uncertainty into an answerable question
2. Systematic retrieval of the best evidence available
3. A critical appraisal of the evidence (e.g. confounding, selection bias etc.)
4. Application of results into clinical practice (see Chapter on Implementation Research)
5. Performance evaluation

Several guidelines have been suggested as a way of assessing the quality of evidence and include the US Preventative Task Force, the UK National Health Service, and the GRADE Working Group. The US Preventive Services Task Force guidelines rank evidence about the effectiveness of treatments or screening (http://en.wikipedia.org/wiki/Levels_of_evidence):

Level I: Evidence obtained from at least one properly designed randomized controlled trial.

Level II-1: Evidence obtained from well-designed controlled trials without randomization.

Level II-2: Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group.

Level II-3: Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled trials might also be regarded as this type of evidence.

Level III: Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

UK National Health Service

The UK National Health Service uses a similar system with categories labeled A, B, C, and D. These levels are only appropriate for treatment or interventions; different types of research are required for assessing diagnostic accuracy or natural history and prognosis, and hence different "levels" are required. For example, the Oxford Centre for Evidence-based Medicine suggests levels of evidence (LOE) according to the study designs and critical appraisal of prevention, diagnosis, prognosis, therapy, and harm studies [34]:

- Level A: Consistent randomised controlled clinical trial, cohort study, all or none (see note below), clinical decision rule validated in different populations.
- Level B: Consistent retrospective cohort, exploratory cohort, ecological study, outcomes research, case-control study; or extrapolations from level A studies.
- Level C: Case-series study or extrapolations from level B studies.
- Level D: Expert opinion without explicit critical appraisal, or based on physiology, bench research or first principles.

Categories of Recommendations

In guidelines and other publications, recommendations for a clinical service are classified by the balance of risk versus benefit of the service *and* the level of evidence on which this information is based. The U.S. Preventive Services Task Force uses [35]:

- Level A: Good scientific evidence suggests that the benefits of the clinical service substantially outweigh the potential risks. Clinicians should discuss the service with eligible patients.
- Level B: At least fair scientific evidence suggests that the benefits of the clinical service outweigh the potential risks. Clinicians should discuss the service with eligible patients.
- Level C: At least fair scientific evidence suggests that there are benefits provided by the clinical service, but the balance between benefits and risks are too close for making general recommendations. Clinicians need not offer it unless there are individual considerations.
- Level D: At least fair scientific evidence suggests that the risks of the clinical service outweigh potential benefits. Clinicians should not routinely offer the service to asymptomatic patients.
- Level I: Scientific evidence is lacking, of poor quality, or conflicting, such that the risk versus benefit balance cannot be assessed. Clinicians should help patients understand the uncertainty surrounding the clinical service.

The Grading of Recommendations Assessment, Development and Evaluation (The GRADE Working Group)

A newer system was developed by the GRADE working group and takes into account more dimensions than just the quality of medical research. It requires users of GRADE who are performing an assessment of the quality of evidence, usually as part of a systematic review, to consider the impact of different factors on their confidence in the results. Authors of GRADE tables, divide the quality of evidence into four levels, on the basis of their confidence in the observed effect (a numerical value) being close to what the true effect is. The confidence value is based on judgments assigned in five different domains in a structured manner. The GRADE working group defines 'quality of evidence' and 'strength of recommendations' as two different concepts which are commonly confused with each other.

Systematic reviews may include randomized controlled trials that have low risk of bias, or, observational studies that have high risk of bias. In the case of randomized controlled trials, the quality of evidence is high, but can be downgraded in five different domains.

- Risk of bias: Is a judgment made on the basis of the chance that bias in included studies has influenced the estimate of effect.
- Imprecision: Is a judgment made on the basis of the chance that the observed estimate of effect could change completely.
- Indirectness: Is a judgment made on the basis of the differences in characteristics of how the study was conducted and how the results are actually going to be applied.
- Inconsistency: Is a judgment made on the basis of the variability of results across the included studies.
- Publication bias: Is a judgment made on the basis of the question whether all the research evidence has been taken to account.

In the case of observational studies, the quality of evidence starts out lower and may be upgraded in three domains in addition to being subject to downgrading.

- Large effect: This is when methodologically strong studies show that the observed effect is so large that the probability of it changing completely is less likely.
- Plausible confounding would change the effect: This is when despite the presence of a possible confounding factor which is expected to reduce the observed effect, the effect estimate still shows significant effect.
- Dose response gradient: This is when the intervention used becomes more effective with increasing dose. This suggests that a further increase will likely bring about more effect.

Meaning of the levels of quality of evidence as per GRADE

- High Quality Evidence: The authors are very confident that the estimate that is presented lies very close to the true value. One could interpret it as: there is very low probability of further research completely changing the presented conclusions.
- Moderate Quality Evidence: The authors are confident that the presented estimate lies close to the true value, but it is also possible that it may be substantially different. One could also interpret it as: further research may completely change the conclusions.
- Low Quality Evidence: The authors are not confident in the effect estimate and the true value may be substantially different. One could interpret it as: further research is likely to change the presented conclusions completely.

- Very Low Quality Evidence: The authors do not have any confidence in the estimate and it is likely that the true value is substantially different from it. One could interpret it as: New research will most probably change the presented conclusions completely.

Guideline panelists may make strong or weak recommendations on the basis of further criteria. Some of the important criteria are:

- Balance between desirable and undesirable effects (not considering cost)
- Quality of the evidence
- Values and preferences
- Costs (resource utilization)

Despite the differences between systems, the purposes are the same: to guide users of clinical research information on which studies are likely to be most valid. However, the individual studies still require careful critical appraisal.

In summary, the term EBM has been linked to three potentially false premises: that evidence has a purely objective meaning in biomedical science; that one can distinguish between what is evidence and what is lack of evidence; and that there is evidence-based, and non-evidence-based medicine. As long as it is remembered that the term evidence, while delivering forceful promises of truth, is limited in the sense that scientific work can never prove anything but only serves to falsify, the term has some usefulness. Finally, EBM does rely upon the ability to perform systematic reviews (meta-analyses) of the available literature, with all the attendant limitations of meta-analyses discussed above.

In a "tongue and cheek" article, Smith and Pell addressed many of the above issues in an article entitled "*Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomized controlled trials*" [37]. In their results section, they note that they were unable to find any RCTs of "parachute intervention". They conclude that:

> *only two options exist. The first is that we accept that under exceptional circumstances, common sense might be applied when considering the potential risks and benefits of interventions. The second is that we continue our quest for the holy grail of exclusively evidence-based interventions and preclude parachute use outside of a properly conducted trial. The dependency we have created in our population may make recruitment of the unenlightened masses to such a trial difficult. If so, we feel assured that those who advocate evidence-based medicine and criticize use of interventions that lack evidence-base will not hesitate to demonstrate their commitment by volunteering for a double blind, randomized, placebo controlled, crossover trail.* (See Fig. 10.3)

Isaacs has embellished this with a list for the basis of clinical decision making (Table 10.11) in which evidence is one, and then eminence, vehemence, eloquence, providence, diffidence, nervousness, and confidence round out the list [38]. For each they describe the bias as follows: eminence based medicine-"the more senior the colleague, the less importance he or she placed on the need for anything as mundane as evidence"; vehemence based medicine- is determined by the loudest colleague; eloquence based medicine is predicted on sartorial elegance as a powerful substitute for evidence; providence based medicine occurs when you have no clue what to

HULTONGETTY

Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomised controlled trials

Fig. 10.3 Humorous example of evidence-based medicine (With permission: Smith and Pell [37])

Table 10.11 Humorous outline of the basis of clinical decision making

Basis for clinical decision making	Marker	Measuring device	Unit of measurement
Evidence	RCT	Meta-analysis	Odds ratio
Eminence	Grey hair	Luminometer	Optical density
Vehemence	Stridency	Audiometer	Decibels
Eloquence	Sartorial splendor	Teflometer	Adhesion score
Providence	Religious fervor	Genuflection angle	Piety units
Diffidence	Gloom level	Nihilometer	Sighs
Nervousness	Litigation phobia	Every conceivable test	Bank balance
Confidence	Bravado	Sweat test	No sweat

Adapted from: Isaacs and Fitzgerald [38]

do and you turn to God to give you a hand with decision making; diffidence based medicine is when nothing is done out of a sense of despair, however they further point out that this may be beneficial since doing something may be worse ("don't just do something, stand there" as the axiom goes); nervousness based medicine is decision making based on fear of litigation (here the only bad test is "the one you didn't think of ordering"); and finally, confidence based medicine, which the authors point out is restricted to surgeons.

Clinical Practice Guidelines

Another outcropping from evidence-based medicine is clinical practice guidelines. Guideline recommendations have become the standard of care, and quality of care is increasingly assessed on the basis of adherence to these recommendations. In 1990 the Institute of Medicine defined practice guidelines as *"systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances*" [39]. In an editorial, Gibbons et al. noted that "*as the number of available guidelines provided by a variety of sources has literally exploded, serious questions and controversies have arisen about how guidelines should be developed, implemented, and evaluated.*" [40] They go on to point out that guideline developers have been criticized for failing to control for conflicts of interest, for variable quality, and for failing to prove that guidelines benefit patients.

Despite the fact that guideline recommendations are being used to asses standard of care, clinical practice guidelines are recommendations (not rules or standards) about the care of patients with a specific condition and ideally are based upon the "best available evidence", but should always be tempered on the basis of individual patient circumstances and preferences. The American Academy of Family Practice defines guidelines as *"a recommendation issued for the purpose of influencing decisions about health interventions*." The "best available evidence" is generally considered evidence from systematic reviews ideally of randomized controlled trials. Although guidelines are intended for clinicians, they are (perhaps unfortunately) used by others to monitor physician practice and in medical-legal proceedings. These aforementioned uses sometimes do not recognize that guidelines are suggestions for care not mandates, and only apply to a percentage of patients with a condition and certainly not all patients. However, the impetus for practice guidelines is many and includes:

- Increasing/changing medical knowledge
- Rising health care costs unrelated to health outcomes
- Wide variations in clinical decisions
- Desire for evidence-based, outcomes-oriented clinical decisions

The reality in medicine is that there has been an explosion of knowledge technology, and of patient expectations and "just keepin' up" with the literature (much less reviewing older literature) is problematic. For example in 1998 there were at least 20,657 articles involving human beings. If one slept 4 h a night, spent 25 h a week seeing patients and 1 h a day on personal activities, and read three articles an hour in the remaining awake time, after 1 year one would be 3,800 additional articles behind. There is no question that the complexity of medical decisions is rapidly growing and that there is uncertainty and variability in medical practice and even the best-trained physician with the greatest experience is not perfect. The above-average physician has even more problems with consistency and accuracy. Thus, there is variability in clinical judgment, a question about the reliability of diagnostic judgment (If a doctor tells you that you have a disease, do you have it? If a doctor does not find a disease, are you well?), and physician decisions can be highly variable (it is well known that physicians can disagree with their peers who have reviewed the same patient, and that they can disagree with themselves when presented with the same patient records at two points in time). An example of this aforementioned disagreement is a study of four cardiologists presented with high-quality angiograms and asked to determine if stenosis in the proximal or distal left anterior descending artery was >50 %.

- The cardiologists disagreed on 60 % of the cases [41]
- Cardiologists looking at the same angiograms at two points in time disagree with themselves 8–37 % of the time [42]

It is also known that there is substantial geographic variability in the rates of procedures.

Guideline recommendations come from medical textbooks, review articles, meta-analyses, expert opinion and consensus panel recommendations, but whereas the US government was once the primary source of guidelines, this is now mostly the province of specialty and subspecialty societies with the exception of the US Preventive Service Task Force. There are instances where there is disagreement amongst the guidelines and although the disagreements are usually minor, the disagreements are certainly a barrier to their acceptance, although clinicians are most likely to accept the recommendations from their own specialty society (and least likely to accept recommendations from managed care organizations or insurance companies).

Some guideline panels use a grading system (discussed above) attached to their recommendations based on the strength of evidence leading to the recommendation.

Summary of Concerns About Guidelines

- Guidelines are often outdated by the time they are released. (Burn your textbooks, except this one, of course)
- Guidelines often emphasize peer consensus rather than outcome evidence
- Guidelines ignore patient preference.

Other Concerns

Evidence-based guidelines disregard effective treatments that have not been evaluated in systematic experimental studies. A treatment might get a low rating because it does not work *or* because it has not been evaluated in a randomized clinical trial. Evidence-based medicine assumes that untested treatments are ineffective. Finally, many clinicians view practice guidelines as “cook book medicine” with “not enough recipes in the cookbook” [43].

The limitations in the evidence EBM is nicely reviewed by Sniderman et al. in response to a commentary by Prasad who discusses the two medical world views of whether RCTs are needed to accept new practices [44]. Some of the limitations discussed by Sniderman et al. include:

For many clinical problems there simply is no RCT evidence
Other times multiple RCTs have been performed but the conclusions are in conflict
RCTs are limited in their generalizability

There are limitations in applying the results in a group of patients to the individual

In an attempt to overcome some of the above limitations meta-analyses are performed, but meta-analyses have their own set of limitations (see above)

There are limitations in the guideline process which is also developed to address some of the above problems (e.g. conflicts of interest, failure to ensure dissenting and minority viewpoints, the absence of a process to challenge the validity of specific conclusions that guidelines reach)

There can be a diminution of clinical reasoning as a result of guideline recommendations

References

1. Meinert CL. Meta-analysis: science or religion? Control Clin Trials. 1989;10:257S–63S.
2. Boden WE. Meta-analysis in clinical trials reporting: has a tool become a weapon? Am J Cardiol. 1992;69:681–6.
3. Oxman AD. Meta-statistics: help or hindrance? ACP J Club. 1993;118:A-1–13.
4. Goodman SN. Have you ever meta-analysis you didn't like? Ann Intern Med. 1991;114:244–6.
5. Bangalore S. Dueling data: separating the wheat from the statistical chaff. CardioSource WorldNews. 2012;Dec:14.
6. Pearson K. Report on certain enteric fever inoculation statistics. Br Med J. 1904;3:1243–6.
7. Beecher HK. The powerful placebo. JAMA. 1955;159:1602–6.
8. Glass G. Primary, secondary and meta-analysis of research. Educ Res. 1976;5:3–8.
9. Petitti DB. Approaches to heterogeneity in meta-analysis. Stat Med. 2001;20:3625–33.
10. Neveol A, Dogan RI, Lu Z. Author keywords in biomedical journal articles. AMIA Ann Symp Proc/AMIA Symp. 2010;2010:537–41.
11. Turner EH, Knoepflmacher D, Shapley L. Publication bias in antipsychotic trials: an analysis of efficacy comparing the published literature to the US Food and Drug Administration database. PLoS Med. 2012;9:e1001189.
12. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002. doi:10.1136/bmj.d4002.
13. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50:1088–101.
14. Egger M, Smith DG, Altman DG. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2000.
15. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56:455–63.
16. Candelise L, Ciccone A. Gangliosides for acute ischaemic stroke. Cochrane Database Syst Rev. 2001;4:CD000094.
17. Easterbrook. Publication bias in clinical research. Lancet. 1991;337:867.
18. Smith ML. Publication bias and meta-analysis. Eval Educ. 1980;4:22–4.
19. Glass G. Meta-analysis at 25. 2000. Available at: http://glass.ed.asu.edu/gene/papers/meta25.html
20. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Br Med J. 2003;327:557–60.
21. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. Br Med J. 2008;336:1413–5.
22. Chalmers TC, Smith Jr H, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials. 1981;2:31–49.

23. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2000.
24. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22:719–48.
25. Berlin JA, Colditz GA. The role of meta-analysis in the regulatory process for foods, drugs, and devices. JAMA. 1999;281:830–4.
26. Stroup DF, Berlin JA, Morton SC, Olkin L, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology. JAMA. 2000;283:2008–12.
27. Carroli A, Mackey ME, Bergel E. Critical appraisal of systematic reviews. The World Health Organization. www.casp-uk.net. Accessed 20 Aug 2013.
28. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312:71–2.
29. The Cochrane Library. Issue 2. Chichester: Wiley; 2007.
30. Evidence-Based Medicine. 1999. http://library.uchc.edu/lippub/fall99.PDF. Accessed 29 July 2013.
31. Panda A, Dorairajan LN, Kumar S. Application of evidence-based urology in improving quality of care. Indian J Urol. 2007;23:91–6. PMC2721549.
32. Uniformed Services University James A Zimble Learning Resource Center. Evidence-Based Medicine (EBM) Resources. 2000; Available from: www.lrc.usuhs.edu/lrcguides/?q=node/16
33. The Problem of Induction. 1953, 1974. Accessed at http://dieoff.org/page126.htm
34. Baum SJ. Evidence-based medicine: what's the evidence? Clin Cardiol. 2012;35:259–60. doi:10.1002/clc.21968.
35. Prasad V, Gall V, Cifu A. The frequency of medical reversal. Arch Intern Med. 2011;171:1675–6. doi:10.1001/archinternmed.2011.295.
36. Cook DJ, Jaeschke R, Guyatt GH. Critical appraisal of therapeutic interventions in the intensive care unit: human monoclonal antibody treatment in sepsis. J Club Hamilt Reg Crit Care Gr J Intensive Care Med. 1992;7:275–82.
37. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. Br Med J. 2003;327:1459–61.
38. Isaacs D, Fitzgerald D. Seven alternatives to evidence based medicine. Br Med J. 1999;319:1618–9.
39. Institute of Medicine. Clinical practice guidelines: directions for a new program, Committee to advise the public health service on clinical practice guidelines. In: Field MJ, Lohr KN, editors. US Dept of Health and Human Services. Washington, DC: National Academy Press; 1990.
40. Gibbons GH, Shurin SB, Mensah GA, Lauer MS. Refocusing the agenda on cardiovascular guidelines: an announcement from the National Heart, Lung, and Blood Institute. Circulation. 2013;128:1713–5. doi:10.1161/CIRCULATIONAHA.113.004587.
41. Zir LM, Miller SW, Dinsmore RE, Gilbert JP, Harthorne JW. Interobserver variability in coronary angiography. Circulation. 1976;53:627–32.
42. Detre KM, Wright E, Murphy ML, Takaro T. Observer agreement in evaluating coronary angiograms. Circulation. 1975;52:979–86.
43. Parmley WW. Clinical practice guidelines. Does the cookbook have enough recipes? JAMA. 1994;272:1374–5.
44. Prasad V. Why randomized controlled trials are needed to accept new practices: 2 medical worldviews. Mayo Clin Proc. 2013;88:1046–50. doi:10.1016/j.mayocp.2013.04.026.

Chapter 11
Research Methods for Genetic Studies

Sadeep Shrestha and Donna K. Arnett

Abstract This chapter introduces the basic concepts of fundamental methods for genotype-phenotype association studies and relevant issues in interpretation of genetic epidemiology studies to clinicians. An overview of genetic association studies is provided, which is the current state-of-the-art for clinical and translational genetics. Discussion of future directions in this field is also included.

Keywords Genetic research • Genomics • Hardy Weinberg disequilibrium • Familial aggregation • Linkage disequilibrium • Genome-wide association

This chapter introduces the basic concepts of genes and genetic studies to clinicians. Some of the relevant methods and issues in genetic epidemiology studies are briefly discussed with an emphasis on association studies which are currently the main focus of clinical and translational genetics.

Genetics is the fundamental basis of any organism so understanding of genetics will provide a powerful means to discover hereditary elements in disease etiology. In recent years, genetic studies have shifted from disorders caused by a single gene (e.g. Huntington's disease) to common multi-factorial disorders (e.g. hypertension) that result from the interactions between inherited gene variants and environmental factors, including chemical, physical, biological, social, infectious, behavioral or nutritional factors.

A new field of science, Genetic Epidemiology emerged in the 1960s as a hybrid of genetics, biostatistics, epidemiology and molecular biology, which has been the major tool in establishing whether a phenotype (any morphologic, biochemical, physiologic or behavioral characteristic or trait of an organism) has a genetic component. A second goal of genetic epidemiology is to measure the relative size of that

S. Shrestha, Ph.D., MHS, M.S. (✉) • D.K. Arnett, Ph.D., M.S.
Department of Epidemiology, School of Public Health,
University of Alabama at Birmingham, Birmingham, AL, USA
e-mail: sshresth@uab.edu; Arnett@uab.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_11,

genetic effect in relation to environmental effects. Morton and Chung defined genetic epidemiology as

> a science that deals with the etiology, distribution, and control of disease in groups of relatives, and with inherited causes of disease in populations [1].

In the era of known human genome sequences from multiple individuals, genetic epidemiology methods have been instrumental in identifying the contribution of genes, the environment, and their interactions to better understand disease processes and biological mechanisms.

Genomic scientists have predicted that comprehensive, genomic-based care will become the norm, with individualized preventive medicine, early detection of illnesses and tailoring of specific treatments to an individual's genetic profile. Practicing physicians and health professionals must be knowledgeable in the principles, applications, and limitations of genetics to understand, prevent, and treat any biological disorders in their everyday practice. The primary objective of any genetic research is to translate information from individual laboratory tests to infer the relevance of segments of the human genome in relation to disease risk. This chapter will focus on the fundamental concepts and principles of genetic epidemiology that are important to help clinicians understand genetic studies.

Important Principles of Genetics

In the nineteenth century, long before DNA was known, an Augustinian clergyman, Gregory Mendel, described genes as the fundamental unit that transmits traits from parents to offspring [2]. Based on the observations from his cross-breeding experiments in his garden, Mendel developed some basic concepts on genetic information which still provides the framework upon which all subsequent work in human genetics has been based. Mendel's first law, referred to as the "*The principle of segregation*", basically states that alleles (alternate forms of the gene or sequence at a particular location of the chromosome) at one of the parent's genes segregate independently of the alleles from another parent. Mendel's law, therefore, states that alleles transmitted to an offspring are random (i.e., a matter of chance). It is now known that segregation of alleles occurs during the process of sex cell formation, known as meiosis. His second law is referred to as "*The principle of independent assortment*" which states that two genetic factors are transmitted independently of one another in the formation of gametes. As a result, new combinations of genes can be present in the offspring that are otherwise not possible in either of the parents. These two principles of inheritance and the concepts of dominance and recessive alleles established the foundation of our modern science of genetics. However, Mendel's law is not always true and there are exceptions to these rules, e.g. loci in the same chromosomes tend to transmit together, a key concept in modern genetic epidemiology.

All human cells except the red blood cells (RBC) have a nucleus that carries the individual's genetic information organized in chromosomes. Chromosomes are composed of molecules called deoxyribonucleic acid (DNA) which contain the basic instructions needed to construct proteins and other cellular molecules. Given the diploid nature, each human inherits one copy of the chromosome from the father and the other from the mother. Humans have 22 pairs of autosomal chromosomes and 2 sex-specific chromosomes (X and Y), where males have XY and females have XX chromosomes.

At the molecular level, DNA is a linear strand of alternating sugars (deoxyribose) and phosphate residues with one of four types of bases attached to the sugar. All information necessary to maintain and propagate life is contained within these four simple bases: adenine (A), guanine (G), thymine (T), and cytosine (C). In addition to this structure of a single strand, the two strands of the DNA molecule are connected by a hydrogen bond between two opposing bases of the two strands (T always bonds with A and C always bonds with G) forming a slightly twisted ladder, also referred as double helix. It was not until 1953 that James Watson and Francis Creek described this structure of DNA which became the foundation for our contemporary understanding of genes and disease.

The basic length unit of the DNA is one nucleotide, or one base pair (bp) which refers to the two bases that connect the two strands. In total, the human DNA contains approximately 3.3 billion base pairs and any two DNA fragments differ only with respect to the order of their bases. Three base units, together with the sugar and phosphate component (referred to as **codons**) translate into amino acids. According to the central dogma of molecular biology, DNA is copied into single stranded ribonucleic acid (RNA) in a process called transcription, which is subsequently translated into proteins. With the knowledge of underlying molecular biology, "*gene*" is defined as the part of the DNA segment that encodes a protein which forms the functional unit of the "hereditary" factor. It is now estimated that there are approximately 27,000 genes. The encoded proteins make intermediate phenotypes which regulate the biology of all diseases, so any difference in the DNA sequence could change the disease phenotype. In many species, only a small fraction of the total sequence of the genome encodes protein, and the function and relevance of the remaining noncoding sequences are still unknown. For example, over 98 % of the human genome is noncoding. However, the Encyclopedia of DNA Elements (ENCODE) project recently reported that over 80 % of DNA in the human genome has some biochemical function, most of which is still unknown. We are still in the infant stage of understanding the significance of the rest of these non-coding DNA sequence; however, the sequence could have structural purposes, or be involved in regulating the use of functional genetic information.

Units of Genetic Measure

Different genetic markers, which are a segment of DNA with a known physical location on a chromosome with identifiable inheritance, can be used as measures for genetic studies. A marker can be a gene, structural polymorphisms (e.g. insertion/deletion) or it can be some section of DNA such as short tandem repeat (STR)

Table 11.1 Some significant DNA sequence variants

Sequence variations	Description
Short Tandem Repeats (STR)	Tandemly repeated simple sequence motifs of 2–7 base lengths
Single Nucleotide Polymorphism (SNP)	Variations in a single nucleotide occurring in >1 % of the population
Structural variants	Variation in the structure of the chromosome, that includes deletions, inversions, rearrangements, copy number variations

and single nucleotide polymorphism (SNP). Recent advancements in molecular technology have resulted in the discovery of numerous DNA markers and the database of each marker is increasing daily. Polymorphism (poly = many and morphism = form) is a DNA sequence variation at any locus (any segment or region in the genome) in the population that has existed for some time and observed in at least 1 % of the population, whereas a mutation is often recent and the frequency in populations is less than 1 %. The terms mutation and polymorphism are often used interchangeably but mostly defined in the context of frequency. Variants within coding regions may change the protein function (missense) or predict premature protein truncation (non-sense) and as a result can have effects ranging from beneficial to mutual to deleterious. Likewise, although introns (intragenic regions between coding sequences) do not encode for proteins, polymorphisms can affect intron splicing or regulation of gene expression. To understand the role of genetic factors with any phenotype, it is important to understand these sequence variations among those with and without the phenotype within (population) and between (family) generations. We briefly describe the commonly used markers for genetic testing (Table 11.1).

Short Tandem Repeats (STRs)

STRs are tandemly repeated simple DNA sequence motifs of 2–7 bases in length that are arranged head-to-tail and are well distributed throughout the human genome, primarily in the intragenic regions. They are abundant in essentially all ethnically and geographically defined populations and are characterized by simple Mendelian inheritance. STR polymorphisms originate due to mutations caused by slipped-strand mispairing during DNA replication that results from either the gain or loss of repeat units. Mutation rates typically range from 10^{-3} to 10^{-5} events per gamete per generation, compared to single nucleotide rates of mutation of 10^{-7} to 10^{-9}. In humans, STR markers are routinely used in gene mapping, paternity testing and forensic analysis, linkage and association studies, along with evolutionary and other family studies. STRs have served as valuable tool for linkage studies of monogenic diseases in pedigrees, but have limited utility for candidate gene association studies.

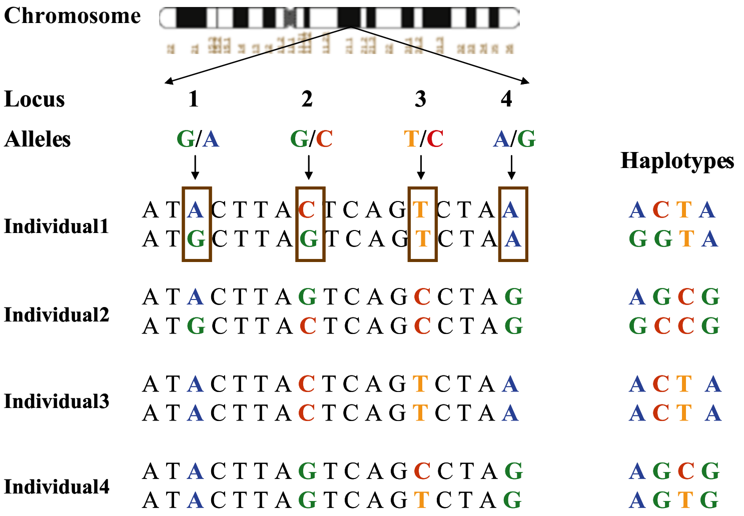

Chromosome
Locus
1
2
3
4
Alleles
G/A
G/C
T/C
A/G
Haplotypes
Individual1
A T A C T T A C T C A G T C T A A
A T G C T T A G T C A G T C T A A
A C T A
G G T A
Individual2
A T A C T T A G T C A G C C T A G
A T G C T T A C T C A G C C T A G
A G C G
G C C G
Individual3
A T A C T T A C T C A G T C T A A
A T A C T T A C T C A G T C T A A
A C T A
A C T A
Individual4
A T A C T T A G T C A G C C T A G
A T A C T T A G T C A G T C T A G
A G C G
A G T G

Fig. 11.1 Alleles and genotypes determined for bi-allelic Single Nucleotide Polymorphisms at four different loci and the corresponding haplotypes. At locus 1, G and A are the alleles; Individuals 1 and 2 have AG heterozygote genotype and Individuals 3 and 4 have AA homozygote genotype. If the phase is known as shown above, the haplotypes for individual 1 would be ACTA and GGTA. However, in most cases, the variant loci are not physically close and the assays may not be able to partition the phase, thus haplotypes are usually estimated with various methods

Single Nucleotide Polymorphisms (SNPs)

SNPs are the variations that occur at a single nucleotide of the sequence. Ninety percent of the polymorphisms in the genome are single nucleotide polymorphisms (SNPs). It has been estimated that there are over 17 million SNPs (1 in every 180 base pairs on average). Most of these variants have been identified through massive efforts of the International HapMap Project (2003) and the 1000 Genomes Project (2008). SNPs are the markers of choice for association studies because of their high frequency, low mutation rates and the availability of high-throughput detection methods. Most SNPs are found in the non-coding region and often have no known biological function, but may be surrogate markers or be involved in regulation of gene (e.g. expression and splicing). With few exceptions, the majority of the SNPs are bi-allelic and the genotypes (genetic makeup at both chromosomes) can be heterozygote (different allele in each chromosome) or homozygote (same allele in both chromosomes) for either allele (Fig. 11.1). All SNPs are catalogued centrally in major databases such as the dbSNP at the National Center for Biotechnology Information (NCBI) and given unique identifiers (rs#) for standard reference.

Structural Variants

The human genome consists of a myriad of structural variants that include deletions, duplications, inversions, translocations and copy number variations (CNVs) that can influence the functions of the encoded proteins. CNVs are the most common structural variants and have been associated several phenotypes and diseases.

It was generally thought that genes occurred in two copies in the genome. Recent studies have suggested that large segments of DNA, ranging from 1 kb to several million bp can vary in copy number, some of which contain several genes. Such CNVs are more common in the human genome than originally thought and can have dramatic phenotypic consequences as a result of altering gene dosage, disrupting coding sequences, or perturbing long-range gene regulation [3]. These regions are estimated to cover 5–20 % of the whole genome.

Although there are different genetic markers (as described above), SNPs are the most frequent variant in the genome and are widely used in genetic studies, so we will refer to SNP polymorphisms to explain the basic concepts in genetic epidemiology, especially in the context of association studies.

Terms and Basic Concepts in Genetic Epidemiology (Table 11.2)

Hardy-Weinberg Equilibrium (HWE)

HWE is one of the key concepts of population genetics that can be used to determine whether a genetic variant could be a valid marker in genetic epidemiology studies. In HWE, allele and genotype frequencies are related through the Hardy-Weinberg law which states that if two alleles, "A" and "a", at any locus with frequencies "p" and "q", respectively, are in equilibrium in a population, the proportions of the genotypes, "AA" homozygotes, "Aa" heterozygotes and "aa" homozygotes will be p^2, 2pq, and q^2, respectively. This law holds as a consequence of random mating in the absence of mutation, migration, natural selection, or random drift. One of the implications of HWE is that the allele frequencies and the genotype frequencies remain constant from generation to generation maintaining equilibrium in overall genetic variations. Extensions of this approach can also be used with multi-allelic and X-linked loci. Deviation from these proportions could indicate (a) genotyping error (b) presence of non-random mating, thus bias in the control selection (c) existence of population stratification (as described later) or (d) recent mutation, migration or genetic drift that has not reached equilibrium. Cases are more likely to represent the tail of a distribution of disease, and any putative genetic variant for that disease may not be in HWE; therefore, it is generally recommended to assess HWE in non-diseased (control) groups.

Table 11.2 Some commonly used genetic terms

Term	Brief description
Hardy-Weinberg Equilibrium (HWE)	Used to determine whether a genetic variant could be a valid marker
Linkage	When two genetic loci are transmitted together from parent to offspring more often than expected
Linkage Disequilibrium (LD)	The extent of non-random association between two genetic loci
Haplotype (Fig. 11.1)	A specific combination of alleles along a chromosome, one from the father and one from the mother
Epigenetic changes	Biochemical alterations in DNA that affect gene expression and function without altering DNA sequence
Transmission Disequilibrium Test (TDT)	Alleles of parents are used as "virtual control" genotypes
LOD score	Logarithm$_{10}$ of odds-the likelihood of observing a segregation pattern of recombination frequency compared to chance

Hardy-Weinberg equilibrium: The stable frequency distribution of genotypes, AA, Aa, and aa, in the proportions p^2, 2pq, and q^2 respectively (where p and q are the frequencies of the alleles, A and a, respectively) that results from random mating in a population in the absence of mutation, migration, natural selection, or random drift

Linkage: co-segregation of alleles at two or more loci (family-based)

Linkage disequilibrium: the extent and associations of non-randomness of alleles at two/more loci in a population

Haplotype: A set of closely linked genetic markers present on one chromosome which tend to be inherited together (e.g. Fig. 11.1 – ACTA and GGTA for individual 1)

Epigenetic Changes: genetic control of the expression and activation of genes that involves factors other than changes in DNA sequence

Transmission Disequilibrium Test (**TDT**): a test that measures overtransmission of alleles from parents to offspring with the disease/trait (more frequently than expected by chance)

LOD Score: Logarithm$_{10}$ odds of likelihood of observing the segregation pattern of the marker alleles at a given recombination frequency (linked) to the likelihood of the same segregation pattern in the absence of linkage (by chance)

Linkage and Linkage Disequilibrium (LD)

Linkage and LD are the *sine qua non* of genetic epidemiology. While genes in different chromosomes segregate, Thomas Hunt Morgan and his co-workers observed that genes physically linked to one another on chromosomes of drosophila tended to be transmitted together. This phenomenon, where two genetic loci are transmitted together from parent to offspring more often than expected under independent inheritance, is termed linkage. Linkage was first demonstrated in humans by Julia Bell and J.B.S Haldane who showed that hemophilia and color blindness tended to be inherited together in some families [4]. Two loci are linked if recombination (exchange of genetic information between two homologous chromosomes during meiosis) occurs between them with a probability of less than 50 %. Recombination is inversely related to the physical distance between the two loci. However, after several generations, successive recombinations (especially in regions of recombination hotspots) may lead to complete independence even between loci that may be physically close together.

In population genetics, LD is defined as the extent of non-random association between two genetic loci such that the presence of one allele at a locus provides information about the allele of the other loci [5]. The level of LD in a population is influenced by several factors including genetic linkage, the rate of recombination, mutation, random genetic drift, selection, non-random mating and population admixture. Many different measures of LD have been proposed in the literature, most of which capture the strength of association between pairs of SNPs. Although concepts of LD date to early 1900s, the first commonly used LD measure, D' was developed by Richard Lewontin in 1964. D' measures the departure from allelic equilibrium between separate loci on the same chromosome that is due to the genetic linkage between them. The other pairwise measure of LD used in association studies is r^2 also denoted as Δ^2.

For two loci with alleles A/a at the first locus and B/b at the second allele, D is estimated as follows:

$$D = p_{AB} - p_A p_B \tag{1}$$

The disadvantage of D is that the range of possible value depends greatly on the marginal allele frequency. D′ is a standardized D coefficient and is estimated as follows:

$$D' = \frac{D}{D_{\max}} \tag{2}$$

If $D > 0$, $D_{\max} = \min\,[P_A(1\text{-}P_B),\ P_B(1\text{-}P_A)]$
If $D < 0$, $D_{\max} = \min[P_A P_B,\ (1\text{-}P_A)(1\text{-}P_B)]$s
and r^2 is the correlation between two loci and is estimated as follows:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b} \tag{3}$$

Both D′ and r^2 range from 0 (no disequilibrium) to 1 (complete disequilibrium), but their interpretation is slightly different. In the case of true SNPs, D′ equals 1 if just two or three of the possible haplotypes are present and is <1 if all four possible haplotypes are present. On the other hand, r^2 is equal to 1 if only two haplotypes are present. Association is best estimated using the r^2 because it acts as a direct correlation to the allele at the other SNP. Additionally, there is a simple inverse relationship between r^2 and the sample size to detect association between susceptibility loci and SNPs.

Haplotype

Haplotype is a specific combination of alleles along a chromosome, one inherited from the mother and the other from the father (Fig. 11.1). Recent studies have shown that the human genome can be parsed into discrete blocks of high LD interspersed by shorter regions of low or no LD. Only a small number of characteristic ("tag") SNPs are sufficient to capture most of the haplotype structure of the human genome in each block. Tag SNPs are loci that can serve as proxies for many other SNPs such that only a subset of loci needs to be genotyped to obtain the same information and power obtained from genotyping a larger number of SNPs. The SNPs within the same block show a strong LD pattern while those in different blocks generally show a weak LD pattern. This advantage, along with the relatively smaller number of haplotypes defined by tag SNPs in each block provides another way to resolve the complexity of haplotypes.

High LD between adjacent SNPs, also result in a much smaller number of haplotypes observed than the theoretical number of all possible haplotypes (2^n haplotypes for n SNPs). There is also biological evidence that several linked variations in a single gene can cause several changes in the final protein product and the joint effect can have an influence on the function, expression and quantity of protein resulting in the phenotype variation. The most robust method to determine haplotypes is either pedigree analysis or DNA sequencing of cloned DNA. Both of these methods are limited by data collection of families or intensive laboratory procedures, but the **phase** (knowledge of the orientation of alleles on a particular transmitted chromosome) of the SNPs in each haplotype can be directly determined. Haplotypes can also be constructed statistically, although constructing haplotypes from unrelated individuals is challenging because the phase is inferred rather than directly measured. Unless all SNPs are homozygous or at most only one heterozygous SNP is observed per individual, haplotypes cannot be discerned. To account for ambiguous haplotypes, several statistical algorithms have been developed [6]. Three common algorithmic approaches used in reconstructing population-based haplotypes are (i) a parsimony algorithm, (ii) a Bayesian population genetic model that uses coalescent theory, and (iii) a maximum likelihood approach that is based on expectation-maximization (EM) algorithm. The details of these methods are beyond the scope of this book, but readers are referred to the book "Computational Methods for SNPs and Haplotype Inference" [6] for further discussion. Recent haplotype estimation methods often use a hybrid approach of EM and Bayesian models.

Biological Specimens

Although the focus of this chapter is not on the laboratory methods of specimen collection, we briefly describe the samples used in clinical studies and their importance. Clinicians deal with different biological organs and tissues in their everyday practice. Most of these however may not be an efficient or convenient source for DNA, the most commonly used resource for genetic studies. Based on factors including cost, convenience for collection and storage, quantity and quality of the source, DNA is commonly extracted from four types of biological specimens: (1) dried blood spots collected in special filter paper (2) whole blood collected in ethylenediaminetetraacetic acid (EDTA) or other anticoagulants such as heparin and acid citrate dextrose (ACD) (3) lymphocytes isolated from whole blood and EBV-transformed for unlimited source of DNA and (4) buccal epithelial cells collected from swabs or mouth-washes (non-invasive and child-friendly). In certain circumstances, samples derived from surgery or other treatment or therapy procedures can also be used for extracting DNA. For instance, formalin-embedded samples of biopsies can be used; however, special laboratory protocols or reagents may be needed (for instance to process the DNA crosslinking).

Ethical, Legal and Social Implications (ELSI)

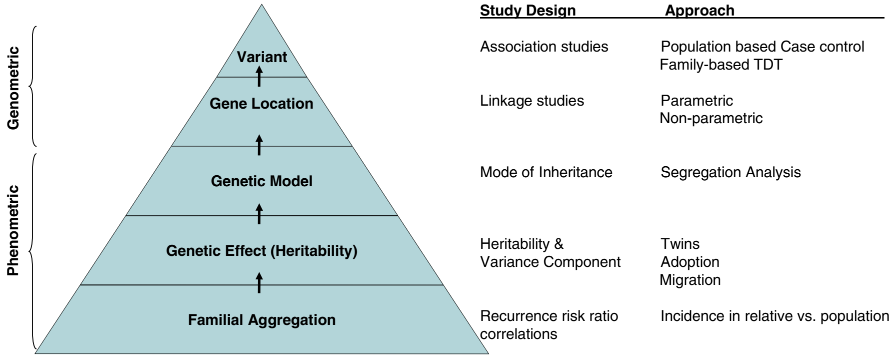
Even for well-intentioned research, one can raise legitimate concerns about the potential misuse of genetic data in regard to social status, employment, economic harm and other factors. A significant amount of work has been done on ethical, legal and social implications (ELSI) research of genetics and policies, but ethics remains an area of major concern. All research protocols can only be conducted upon approval from an institutional review board (IRB) with an appropriate informed consent from the participants. Pediatric genetic research often can be cumbersome as it may require approval from both parents or the legal guardians. It is a routine practice to label the samples with unlinked coded identifiers rather than personal identifiers, so that the individual's identity is masked when linking to phenotypic, demographic, or other personal information. The confidentiality of the DNA results needs to be maximized to protect individual privacy. All reports of genetic studies including manuscripts and grants often require detailed description of ethical concerns and data protection.

Measurable Outcome and Phenotype

Phenotype is an observable and measurable trait which can be defined qualitatively or quantitatively and does not necessarily have to be related to a disease. Some traits or diseases, like the simple Mendelian traits, have a distinctly measurable phenotype definition. However, other illnesses (e.g. psychiatric disorders) are complex to define and require various symptoms and clinical criteria that may have different biological system and pathways combined. The misclassification of cases and controls can be a major problem in any study that can easily introduce biases and inconsistencies between studies. Phenotypes can be defined qualitatively (absent or present) or measured quantitatively. A qualitative trait can be categorized into two or more groups. For example, qualitative traits can be dichotomous (e.g. HIV^+ vs. HIV^-), ordinal (low, average and high blood pressure group) or nominal (green, black, blue eyes) based on certain distinct criteria. On the other hand, measurable physiological quantities such as height, blood pressure, serum cholesterol levels, and body mass index (BMI) can vary among different individuals. Often it may be difficult to examine the genetic effect of quantitative measures; however, they can be transformed into meaningful qualitative values where the genetic effect can be more distinct. To make the quantitative traits more interpretable through statistical analyses, the overall distribution in a given population is viewed graphically. Often these distributions produce a familiar bell-shaped curve (normal distribution), where several statistical methods can be used to assess the effect of genotypes. For example, the individuals at the extreme tails of the curves can have different genetic distributions. Some diseases may also have intermediate phenotypes that can be measured with molecular markers, while others are strictly based on clinical diagnoses. For example, blood cholesterol levels which can be precisely measured may be a better intermediate outcome of cardiovascular disease than a self reported "headache" where the symptoms may be heterogeneous in the population and the measurement is subjective. Other measures, including exposures (e.g. HIV viral load) can define a phenotype better than the clinical symptoms since virally infected individuals can be asymptomatic for undefined period of time. In that specific example, everyone positive for HIV virus test could be defined as the outcome of interest (cases) while in another scenario specific clinical symptoms of HIV infection (e.g. immune cell counts or viral load) could define case status. Even among phenotypes with clinical diagnoses, some have distinct symptoms or signs, with high sensitivity tests, whereas others do not. Some diseases, like Alzheimer's, can have phenotypic heterogeneity, where the same disease shows different features in different families or subgroups of patients. Like in any other clinical study, the key to a genetic study is a clear and consistent definition of the phenotype with underlying biology. Since the main interest in conducting genetic study is to see how variants that change the expression and encoding of protein is related to the biology of the disease, the phenotype has to be clearly defined.

General Methods in Clinical Genetic and Genetic Epidemiology Studies

In the past 20–30 years, epidemiologic methods and approaches have been integrated with those of basic genetics to identify the role of genetic factors in disease occurrence in families and populations [7]. Family studies examine the rates of diseases in the relatives of proband cases versus the relatives of internally matched controls. For a quantitative trait, such as blood pressure, we can measure correlation of trait values among family members to derive estimates of heritability. Mendelian diseases are transmitted in families and recur in the relatives of affected individuals more frequently (10^3–10^6 fold) compared to the general population. In contrast, diseases such as cancer, Alzheimer's Disease, and myocardial infarction are quite common among older adults; however, their occurrence does not follow Mendelian inheritance patterns, but rather are multifactorial with several interactions between environment and genetic factors.



Study Design	Approach
Association studies	Population based Case control Family-based TDT
Linkage studies	Parametric Non-parametric
Mode of Inheritance	Segregation Analysis
Heritability & Variance Component	Twins Adoption Migration
Recurrence risk ratio correlations	Incidence in relative vs. population

Fig. 11.2 Systematic designs and approaches in genetic epidemiology studies to identify the genetic and non-genetic causes of disease

The first step in clinical or epidemiologic genetic studies is to determine whether a phenotype of interest is controlled by a genetic component. There are five key scientific questions that are addressed in sequence in genetic epidemiologic studies (Fig. 11.2): (1) Is there familial clustering? (2) Is there evidence of genetic effect? (3) Is there evidence for a particular genetic model? (4) Where is the disease gene? (5) How does this gene contribute to disease in the general population? The first three questions do not require DNA data and are referred as phenometric studies, but the latter two depend on DNA and referred as genometric studies.

Familial Aggregation

The first step to determine whether a phenotype has a genetic component is to examine the clustering within families. Familial aggregation estimates the likelihood of a phenotype in close relatives of cases compared to the non-cases. If the phenotype is a binary trait, familial aggregation is often measured by the relative recurrence risk. The recurrence risk ratio is the ratio of prevalence of the phenotype in relatives of affected cases to the general population. Greater risk associated with closer degrees of relatedness could also indicate the genetic component. If the prevalence of the phenotype is higher in 1st degree relatives (father, mother, siblings) versus 2nd degree relatives (uncle aunt, cousins) it would suggest a genetic component since the 1st degree relatives share more genetic information than the 2nd degree relatives. For example, cancer and heart disease tend to run in families, as measured by measurements such as relative risk. On the other hand, assessment of familial aggregation of a continuous trait, such as height, can be estimated with a correlation or covariance-based measure such as intrafamily correlation coefficient (ICC). The ICC indicates the proportion of the total variability in a phenotype that can reasonably be attributed to real variability between families. Disease or traits may cluster in families; however, this does not necessarily mean that they share the common genetic factors. Since families often share the same household or geographic region they share common cultural attitudes, socioeconomic status, diet and environmental exposures – all of which can be known or unknown and may not be easily measured. It is difficult to disentangle the genetic effect from the environmental effect due to this shared physical environment. For example, obesity could be due to shared genes within the family or the eating or physical activity habits in the family.

Genetic Effect

Once the familial aggregation is established, the next step is to distinguish between genetic and non-genetic factors and estimate the extent of genetic effect. Different variance component models estimate heritability, which is defined as the proportion of variation directly attributable to genetic differences among relatives to the total variation in the population (both genetic and environmental). Although traditionally used to estimate the genetic effect in familial aggregation, it is a theoretical concept. Heritability is population-specific and must be used with caution when comparing different populations. Other classical designs for distinguishing non-genetic family effects from genetic effects have been studies of twins, adoptees and migrants.

Twin studies

Studies of twins are useful in estimating the contribution to a phenotype through the comparison of monozygotic (MZ) pairs (who share all genes) with dizygotic (DZ) pairs (who share on average half of their genes). If family upbringing acts equally on monozygotic twins as it does on dizygotic twins, then the greater similarity of phenotypes in MZ than DZ twins is attributed to genetic factors. While MZ twins reared together have the same genetic and environment exposures, MZ twins separated at birth and raised apart will have different environment exposures but same genetics. Thus, such studies will provide insights into the contribution of strong environment factors in common diseases such as substance abuse and eating disorders. In contrast, DZ twins may have a similar genetic makeup as other siblings, but they share the same womb, so early environmental exposure related studies can be conducted with these pairs. Concordance rates are used in twin studies which measures and compares the frequency of disease occurrence between MZ and DZ twins. For example, the concordance rate of sickle cell disease among MZ is 100 %, indicating pure genetic effect; whereas Type I Diabetes is 25–35 % among MZ, 5–6 % among DZ twins or siblings and 0.4 % among the general population, suggesting both genetic and environment effects.

Adoption Studies

This study design examines the similarity and differences in the phenotype in the biological parents and foster parents of adoptees, and in their biological and adopted siblings, respectively. The assumptions are that the similarity between an adopted child and biological parent is primarily due to genetic effects, while the similarity between the adopted child and the adoptive parent or adoptive siblings is mainly due to the shared environment since they do not share genetic background as they are not biologically related.

Migration Studies

While with modern globalization, humans are constantly travelling, we are also moving to new areas in search of better opportunities. Patterns in environmental exposures in different areas among different ethnic groups or related family members can be assessed to make some inferences about genetic and environmental influence in phenotypes or diseases. A similar incidence of phenotype or disease in migrants compared to the aboriginal population's incidence suggests a strong environmental factor, whereas similar incidence to the original ethnic group or relatives in the original residence could suggest a genetic effect. Genes do not change as easily as environmental exposures, so the variation in the phenotype after taking into account all the common and new environmental factors could point to a genetic effect.

Genetic Model

After the genetic basis is established, the next step is to find the mode of inheritance which historically was done using segregation analyses, although these methods are not as common in the era of SNP association studies. Segregation analyses does not use DNA-based genetic data, but rather, the methods test whether or not the observed phenotype follows a Mendelian inheritance in the offspring in the pedigree. Mendelian diseases can be autosomal dominant, autosomal recessive, X-linked dominant, or X-linked recessive (usually with high penetrance and low frequency of risk alleles). Traditional segregation analysis primarily studied simple Mendelian disorders where a single gene mutation is sufficient and necessary to cause a disorder. However, most common chronic diseases are regarded as complex where a large number of genetic variants along with environmental factors interact with each other (necessary or un-necessary but not sufficient) to affect the disease outcomes. These diseases usually cluster in families, but do not follow a traditional Mendelian inheritance pattern. While segregation analyses are powerful to test different modes of Mendelian inheritance in the family, it is not useful for complex traits. Linkage and association analysis, both of which utilize DNA, are more powerful to study genetic effects of complex diseases.

Disease Gene Location

Linkage Studies

Linkage studies are performed based on the principle that alleles at two nearby loci on the genome tend to be transmitted together from parent to offspring. Linkage analysis are often the first stage in genetic epidemiology studies to identify broad genomic regions that contain gene or genes that predispose to the phenotype, in the absence of previous biologically driven hypotheses. Genetic linkage analysis tests whether the marker segregates with the disease in pedigrees with multiple affected individuals, according to a Mendelian mode of inheritance. The approach relies entirely on the tendency for genomic regions that affect the phenotype to be passed on to the next generation intact, without recombination events at meiosis. If a marker is passed down through family generation and occurs more commonly among those with the phenotype, then the marker can be used as a surrogate for the location of the gene.

Two types of linkage analysis can be performed: parametric and nonparametric analysis. Parametric linkage analysis involves testing whether the inheritance patterns fits a specific model and is traditionally measured with a statistical test, LOD score (logarithm (base 10) of odds) – $L(\theta)/L(\theta=0.5)$ i.e., the likelihood of observing the segregation pattern of the marker alleles at a given recombination frequency θ (linked) compared with the likelihood of the same segregation pattern in the absence of linkage (by chance). While the approach is very powerful, the study design can be challenging logistically since as it requires recruitment of families (with history of the phenotype) to estimate a number of recombination occurrences in order to calculate the LOD score. STRs with multiple alleles are more powerful for linkage studies than SNPs, which are mostly biallelic. The objective of parametric linkage analysis is to estimate the recombination frequency (θ) and to test whether θ is less than 0.5, which is the case when two loci are genetically linked. The nonparametric approach evaluates the statistical significance of excess allele sharing for specific markers among affected sibs and does not require information about the mode of disease inheritance. With this approach, often the inheritance pattern is measured in terms of identical by descent (IBD), where the same allele is inherited from a common ancestor, and identical by state (IBS), where the allele is the same but not necessarily inherited from the same ancestor. Thus, these methods are based the fact that affected relatives have a higher probability of sharing genes IBD at or near a locus of susceptibility allele/gene to a disease than sharing an unlinked locus. The genes contributing to the phenotypic variation have been successfully localized by linkage analysis for Mendelian diseases that have a strong genetic effect and are relatively rare (e.g. cystic fibrosis, Huntington disease). However, for more complex and common diseases (e.g. cancer, cardiovascular diseases), linkage analysis has had less success. The method of choice for complex genetic diseases has evolved to association studies which are followed by fine-mapping studies to narrow down the putative disease locus.

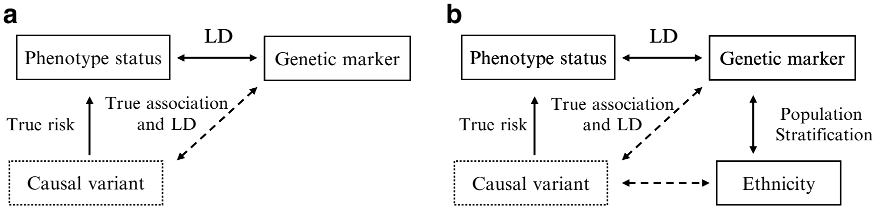


Fig. 11.3 True association, LD and the effect of population stratification. (**a**) Genetic marker that is in LD with causal variant serves as a surrogate of the true association with the phenotype. (**b**) Population stratification is a confounder that leads to spurious association

Association Studies

Genetic association studies aim to correlate differences in allelic frequencies at any locus with differences in disease frequencies or quantitative traits [8]. Genetic association occurs if the specific genetic variant is more frequent in the affected group than the non-affected group. Most association studies represent classical case–control approaches where the risk factor under investigation is the allele at the genetic marker (mostly with SNPs). SNP-based association studies can be performed in two ways: (i) direct testing of an exposure SNP with a known varying function such as altered protein level or structures and (ii) indirect testing of a SNP which is a surrogate marker for locating adjacent functional variant that contributes to the phenotype or disease state (Fig. 11.3a). The first method requires the identification of all "functional" variants in coding and regulatory regions of genes. The latter method avoids the need for cataloguing potential susceptibility variants by relying instead on association between disease and neutral polymorphisms tagging a SNP near a risk-conferring variant. It exploits the phenomenon of linkage disequilibrium (LD) between alleles of closely linked loci within the genomic regions.

Given the diallelic nature of majority of the SNPs, a disease locus may be difficult to identify unless the surrogate marker is closely linked to the disease locus. Apart from a single SNP association strategy, a dense panel of SNPs from the coding and non-coding regions of the gene that form haplotypes can also be tested. Some studies have also demonstrated that the analysis of haplotypes rather than individual SNPs can detect association with complex diseases. It has been suggested that single SNP-based candidate gene studies may be statistically weak as true associations may be missed because of the incomplete information from individual SNPs. For example, haplotypes contain more heterozygosity than any of the individual markers that comprise them and also mark more of the variation in the gene than single SNPs. Several haplotype association studies have shown the power of haplotypes over individual SNPs as it can either combine multiple causal variants or tag a less common causal variant than a more frequent single SNP.

Candidate Gene vs. Genome Wide Association Studies (GWAS)

Candidate gene approaches examine polymorphisms in genes with potential biological mechanisms or pathways related to the phenotype of interest. Some of the candidate genes are also based on physical location or sequence homology to a gene encoding protein that is in the etiologic pathway. As attractive as this hypothesis-driven candidate gene approach is, it focuses exclusively on the relatively few known genes, ignoring many that have not yet been characterized to play a role, suffering from potential publication bias in the process of selection of the genes. One major drawback of candidate gene approach is that *a priori* knowledge of the pathogenesis of the disease is required – when the molecular mechanism is poorly understood or complex, it could lead to selection of the wrong genes. Even with the right genes within the pathway, the challenge is to find variants that influence the regulation of gene function. Candidate gene studies have proven to be more successful when used as a follow-up of linkage studies. For example, APOE4, the most common genetic factor associated with Alzheimer's disease, was primarily discovered by candidate gene approach following the linkage study which mapped to chromosome 19.

Alternatively, with assurance of adequate power, hypothesis-generating genome wide association studies (GWASs) have been widely used. While the study design and methodological approaches are the same as for the candidate gene approach [8], GWAS studies rely on the microarray chips that consists of thousands to millions of genomic variants that has resulted from large projects such as the HapMap, 1000 Genome Project, and continuing sequencing efforts by various groups and investigators. Technological advances have dramatically resulted in cost-effective high-throughput genotyping arrays making GWAS more promising and attractive. GWAS has the advantage in the sense that no *a priori* knowledge of the structure or function of the genes involved is required. Additionally, with complex statistical models, untyped SNPs can also be imputed using GWAS data, which has been proven to be very reliable for common variants. Hence, this approach provides the possibility of identifying variants and genes that influence the phenotype or the disease that had previously not been biologically suspected. A two step design has often been used by researchers where common variation is first screened for association signals using cost-effective typing of tagging SNPs with GWAS followed by denser sets of SNPs in regions of potentially positive signals. If the sample size is large enough, a third stage of validation of association can also be conducted with proper power calculations. Although promising results have been found for different phenotypes with GWAS, analytical considerations are still underway to develop a robust strategy to interpret the findings especially for complex diseases with multiple gene-gene and gene-environmental interactions. Such large datasets still require new methods and approaches to understand the true biology of the phenotype. A lot of emphasis has been made towards using stringent statistical criteria for handling false positive issues; however, new biologically-driven methods are required to dissect such large datasets to understand and identify the complex nature of common diseases.

Risk Quantification

Gene-Gene and Gene-Environment Interaction

A central theme of genetic epidemiology is that human disease is caused by interactions within and between genetic and non-genetic environmental factors. Thus, in the design and analysis of epidemiologic studies, such interaction needs to be explicitly considered. A simple approach would be to create a classic 2×2 table with genotypes at the two loci classified as present or absent and compute odds ratios for all groups with one reference group. The extent of the joint effect of two loci can be compared with the effects for each locus independently. The same approach can be considered for gene-environmental interaction for qualitative measurements. However, as more genes are involved and the environmental exposure is quantitatively measured, the analysis and interpretation of the interaction can be complicated, but various methods are being continuously developed. Large sample sizes are needed to observe true interactions, especially if they are small effects.

Gene Contribution

Once the association of the genetic allele is discovered, it is important to assess the contribution of this variant to the phenotype. The public health relevance of a given polymorphism is addressed by estimating the proportion of diseased individuals in the population that could be prevented if the high-risk alleles were absent (known as attributable fraction, etiologic fraction, or population attributable risk percent). Accurate estimation of the population frequency of the high-risk variant (allele and/or genotype) is important because the **attributable fraction** is a function of the frequency of the high-risk variant in the population and the penetrance (i.e., the likelihood that the trait will be expressed if the patient carries the high-risk variant). Attributable fractions can also be used to estimate the proportion of disease that is a result of the interaction of a genetic variant and an environmental exposure. Genetic variants are not usually modifiable within the longevity of an individual (although very possible evolutionarily over time in populations); therefore the prevention of disease will depend on interventions that target environmental factors that interact with genetic susceptibility to influence the risk of disease.

Additional Applications of Genetic Studies

Most of the genetic studies (candidate or genome-wide) are focused on case–control designs with the underlying goal of understanding the biological cause of the disease. Other time dependent studies can be performed to understand the genetic effect in the natural history or progression of the disease. The outcomes of these studies are helpful for providing counseling to individuals about their offspring (genetic screening) or the interaction between environmental factors. However, there are a growing number of genetic studies examining the differential response to drugs or vaccines, with potential application of translational science. For instance, "**pharmacogenetic**" studies focus on genetic determinants of individual variation in response to drugs, including variation in the primary domain of drug action and variation in risk for rare or unexpected side effects of drugs. Likewise, "**vaccinogenetic**" studies examine the genetic determinants of differential vaccine response (e.g. antibody titer) and side effects between individuals.

Beyond Association Studies

While other factors such as epigenetic and regulatory factors are beyond the scope of this chapter, it is important to understand that association studies itself may not fully delineate the genetic effect on a disease. Epigenetic changes are biochemical alterations in DNA that affect gene expression and function without altering the underlying DNA sequence. DNA methylation is one epigenetic process implicated in human disease that involves methylation of cytosine, usually at CpG dinucleotides. Micro-array methods are available to capture the methylation patterns across genes that could help in addition to the variant findings. Recent insights of the ENCODE project has helped shift focus to complex molecular mechanisms by which genetic factors such as microRNAs (miRNAs) may regulate genes. MiRNAs are evolutionarily conserved small non-coding RNAs (~22 bp) that inhibit translation of proteins by binding to the target transcript in the 3′ untranslated region. It has been estimated that miRNAs contribute to expression of over 60 % of protein coding genes in humans. In this regard, as the testing costs are being lowered, it may be beneficial to perform whole genome sequencing (versus GWAS or even exome-sequencing that targets variants in the exons of all known genes) that will provide information on all the known and the unknown variants of the human genome. Although new approaches and analytical methods are warranted to fully understand the genome, sequencing data will provide both rare and common variants in both genic and non-genic regions which can have regulatory or unknown functions, as suggested by ENCODE.

Major Issues and Limitations in Genetic Studies

In most cases with complex diseases, the effect of any genetic variant is small and can only be observed in studies with a large sample size or the frequency of the allele is rare and has a large relative risk. There are very few common variants (>10 % allele frequency) with a relative risk exceeding 2 (e.g. APOE and Alzheimer's disease). A major concern with respect to genetic association studies has been lack of replication, especially contradictory findings across studies. Replication of findings is very important before any causal inference can be drawn. For example, since 2005, over 1,600 publications have identified more than 2,000 genetic associations with approximately 300 common diseases and traits, but many of these studies need to be replicated. Several study design and statistical issues need to be seriously considered when conducting genetic studies which are briefly described below:

Genetic Heterogeneity

There are several cases where multiple alleles at a locus are associated with the same disease. This phenomenon is known as **allelic heterogeneity** and can be observed with a multi-allelic locus. This may explain why in some studies one allele is associated with the disease and in other studies it is another allele. Likewise, locus heterogeneity may also exist where multiple genes influence the disease independently and thus a gene found to be associated in one study may not be replicated in the other but rather another gene may be associated.

Confounding

One crucial consideration in genetic studies is the choice of an appropriate comparison group. In general, as in any well-designed epidemiological case–control study, controls need to be sampled from the same source population as the cases. The use of convenient comparison groups without proper ascertainment criteria may lead to spurious findings as a result of confounding caused by unmeasured genetic and environmental factors. Population stratification can occur if cases and controls are not matched by ethnicity or if individuals have differential admixture (the proportions of the genome that have ancestry from each subpopulation). Stratification can results when phenotypes of interest differ between ethnic groups (Fig. 11.3b). Although most genetic variation is inter-individual, there is also significant inter-ethnic variation irrespective of disease status. One classic example is reported by Knowler et al. [9] who showed spurious inverse association between variants in the immunoglobulin haplotype Gm3;5,13,14 and non-insulin dependent diabetes mellitus among the Pima-Papago Indians [9]. Individuals with the haplotype Gm3;5,13,14 had a higher prevalence of diabetes than those without it (29 % vs.8 %). This haplotype, however, measured the subjects' degree of Caucasian genetic heritage and when the analysis was stratified by degree of admixture, the association did not exist.

One way to overcome such issue of confounding by population stratification is to conduct family based designs with special statistical analyses such as transmission-disequilibrium test (TDT). Basically, in TDT, alleles of parents not transmitted to the patients are used as “virtual control” genotypes so any population-level allele frequency differences become irrelevant. Several other family-based and population-based methods have also been derived from TDT. While these methods are attractive because they correct false positives from population stratification, family-based samples are difficult to collect and might not be feasible for late-onset diseases where the parents might be deceased. Another approach is to use a “homogeneous” population. In recent years, there is growing interest to study genetically isolated populations such as Finland and Iceland. These populations have been isolated for several years and expanded from a small group of individuals called “**founder population**”. Founder population limits the degree of genetic diversity making more or less a homogenous population. One major limitation of finding from such isolated population is the generalizability to other populations which may have different genetic make-ups.

Studies have shown that there is admixture even within such isolated populations. An alternate method to control for population stratification is to use unrelated markers from the non-functional region of the genome as indicators of the amount of background diversity in individuals. The first approach, referred as “**genomic control**”, measures the extent of inflation due to population stratification and this value can be adjusted in the standard analyses. The second approach would be inferring genetic ancestry, by either the structured-association approach where individuals are assigned to subpopulation clusters using model-based clustering program such as STRUCTURE; or infer population structure with principal component analysis (PCA). Either association analyses are performed by stratifying clusters or covariates derived from ancestry information are adjusted in the analyses.

Genotype Error and Misclassification

For family-based studies (trio data for TDT), genotyping errors have been shown to increase type I and type II errors and for population-based (case–control) studies it can increase type II errors and thus decrease the power. Additionally, misclassification of genotypes can also bias LD measurements.

In general, genotyping errors could be a result of poor amplification, assay failure, DNA quality and quantity, genomic duplication or sample contamination. It is important that a quality-check be performed for each marker and the low-performance once be removed from the analysis before the results are interpreted. Several laboratory based methods such as (a) genotyping duplicate individuals (b) genotyping the same individuals for the same marker using different assay platforms or (c) genotyping in family pedigrees to check for Mendelian inconsistency, (i.e. the offspring should share the genetic makeup of the parents and any deviation could indicate genotype error) can be used to assure the quality of the genotypic data. Testing for HWE is also commonly used, however it is important to note that deviation from HWE does not necessarily indicate genotype error and could be due to any of the underlying causes as described earlier.

Multiple Testing

Regardless of whether each SNP is analyzed one at a time or as part of a haplotype, the number of individual tests can become very large and can lead to an inflated (false positive) type I error rate both in candidate gene approach and whole genome approach. If the selected SNPs are all independent, then adjustments to the conventional p-value of 0.05 with Bonferroni correction could account for the multiple testing. However, given the known LD pattern between SNPs, such adjustments would overcorrect for the inflated false-positive rate, resulting in a reduction in power. An alternate method would be to use the False Discovery Rate (FDR) approach which rather than correcting the p-value, corrects for fraction of false-positives with the significant p-value. When a well defined statistical test is performed (testing a null against an alternative hypothesis) multiple times, the FDR estimates the expected proportion of false positives from among the tests declared significant. For example, if 100 SNPs are said to be significantly associated with a trait at a false discovery rate of 5 %, then on average 5 are expected to be false positives. However, the gold standard approach that is being appreciated more is the permutation testing where the groups status of the individuals are randomly permuted and the analysis repeated several times to get a distribution for the test statistics under the null hypothesis but this method can also be computationally intensive and time-consuming.

Concluding Remarks

The completion of the Human Genome Project in 2003 heightened expectations of the health benefits from genetic studies [10]. Other projects such as the HapMap and 1000 Genome projects have complemented knowledge from the Human Genome Project. The markedly low cost to sequence the genome has provided additional information from various projects, which was not possible a few years ago. The ENCODE project has furthered our knowledge that previously thought “junk” DNA sequences are important as they have regulatory and other unknown functions. While the known genetic factors and methods drive our paths ahead, all the unknown factors make us all strive to answer the multitude of important translational questions in the field of clinical research and medicine.

Methods in genetic epidemiology are very powerful in examining and identifying the underlying genetic basis of any phenotype if conducted properly. There are several study designs that can be used with a common goal of finding both the individual effects and interactions within and between genes and environmental exposures that causes the disease. While the technology has provided us better and efficient platforms to conduct the studies, the underlying purpose of genetic epidemiology studies have always remained the same – what genetic variants cause the phenotype or the disease and how can we complement this deficit or control the

Table 11.3 Possible explanations to consider before interpreting the association study results

Outcomes of association studies	Possible explanations to consider
Positive association	True causal association LD with causal variant confounding by population stratification Hardy Weinberg disequilibrium Multiple comparison (false positive)
Negative association	No causal association Small sample size Phenotype misclassification
Multiple genes associated to the same phenotype	Genetic heterogeneity Interactions within and between genes and environmental factor False positive
Multiple alleles at the same gene associated to the same phenotype	Allelic heterogeneity False positive
Same allele in the same gene associated with the same phenotype but in opposite direction	Confounding by population stratification Phenotype heterogeneity False positive

overload of the protein encoded by the variant in the gene to stop the disease? Regardless of the approach, several design and methodological issues need to be considered when conducting studies and interpreting the results (Table 11.3). Although these studies may find association of the phenotype with a genetic variant, the challenge is to meaningfully translate the findings. In most instances the alleles are in the non-coding region and the frequencies are rare but this the stepping stone in the process of understanding the complexity of common diseases. Very rarely can we find a conclusive evidence of genetic effect from a single study, so replication studies with larger samples size should be encouraged to provide insurance against the unknown confounders and biases. To understand the biologic significance of the variants, animal studies and gene expression studies can be conducted as follow-up studies. Of note, most of the loci from the association studies, singly or in aggregate, only explain a small proportion of trait heritability. This “missing heritability” is reflected by small odds ratios and often has limited predictive utility. Overall, clinicians need to be aware of the potential role of genetics in disease etiology and be cautiously familiar with issues and limitations in conducting genetic epidemiology studies before interpreting them for clinical or public health use.

References

1. Morton NE, Chung CS. Genetic epidemiology. New York: Academic; 1978.
2. Mendel G. Versuche über Pflanzen-Hybriden. Verh. Naturforsch. Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32)1866.

3. Kehrer-sawatzki H, Cooper DN. Copy number variation and disease. Basel/London: S Krager; 2009.
4. Bell J, Haldane JBS. The linkage between the genes for colour-blindness and haemophilia in man. Proc R Soc (Lond) B. 1937;123:119–50.
5. Hartl DL, Clark AG. Principles of population genetics. Sunderland: Sinauer Associates; 2007.
6. Istrail S, Waterman M, Clark A. Computational methods for SNPs and haplotype inference. New York: Springer; 2004.
7. Khoury MJ, Beaty TH, Cohen BH. Fundamental of genetic epidemiology. New York: Oxford University Press; 1993.
8. Ziegler A, Konig IR. Statistical approach to genetic epidemiology: concepts and applications. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA; 2006.
9. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet. 1988;43(4):520–6.
10. Khoury MJ, Burke W, Thomson EJ. Genetics and public health in the 20th century. New York: Oxford University Press; 2000.

Additional References and Recommended Readings

1000 Genomes: http://www.1000genomes.org/
dbSNP National Center for Biotechnology Information (NCBI): http://www.ncbi.nlm.nih.gov/SNP/
ENCODE: https://genome.ucsc.edu/encode/
Human Genome Project: http://www.genome.gov/10001772
International HapMap Project: http://hapmap.ncbi.nlm.nih.gov/

Chapter 12
Research Methods for Pharmacoepidemiology Studies

Maribel Salas and Bruno Stricker

Abstract Pharmacoepidemiology (PE) is the discipline that studies the frequency and distribution of health and disease in human populations, as a result of the use and effects (beneficial and adverse) of drugs. PE uses methods similar to traditional epidemiologic investigation, but applies them to the area of clinical pharmacology. This chapter will review the factors involved in the selection of the type of pharmacoepidemiologic study design, and advantages and disadvantages of these designs. Since other chapters describe randomized clinical trials in detail, we will focus on observational studies.

Keywords Pharmacoepidemiology • Effectiveness trials • Pragmatic trials • Case-time control study

Pharmacoepidemiology (PE) is the discipline that studies the frequency and distribution of health and disease in human populations, as a result of the use and effects (beneficial and adverse) of drugs. PE uses methods similar to traditional epidemiologic investigation, but applies them to the area of clinical pharmacology [1]. Many of the same precepts hold for PE studies as has been discussed in previous chapters, however, this chapter can serve as a review of many of the same principles; but, then as they specifically apply to PE research.

M. Salas, M.D., D.Sc., M.Sc. (✉)
Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania,
Philadelphia, PA, USA

Worldwide Safety Strategy, Pfizer Inc, Collegeville, PA 19426, USA
e-mail: Maribel.Salas@pfizer.com

B. Stricker, M.D., Ph.D.
Department of Epidemiology and Biostatistics, Erasmus University Medical School,
Rotterdam, The Netherlands

Drug Safety Unit, Inspectorate for Health Care, The Hague, The Netherlands

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_12,

In the last few years, PE has acquired relevance because of various drug withdrawals from the market; and, as a result of public scandals related to drug safety and regulatory issues. Some of these withdrawn and controversial drugs include troglitazone, [2–4] cisapride, [5, 6] cerivastatin, [7–10] rofecoxib, [11–13] and valdecoxib [13–15]. One of the major allegations cited with each of these drug withdrawals were flaws in the study designs that were used to demonstrate drug efficacy or safety. Furthermore, the study designs involved with these withdrawn drugs were variable and reported conflicting results [16]. An example of the controversies surrounding drug withdrawals is the association of nonsteroidal antiinflamatory drugs (NSAID) with chronic renal disease [17–21]. The observation that one study may produce different results from another, presumably similar study (and certainly from studies of differing designs) is, of course, not unique to PE, as has been discussed in prior chapters.

Pharmacoepidemiologic studies have been used with many purposes, for example to: examine the natural history of a disease, determine the incidence rates of events in the general population, characterize safety signals associated with medications, describe drug utilization patterns, determine risk factors for specific events, assessing the benefits and risks of products or evaluating strategies to enhance the benefit/risk balance [22].

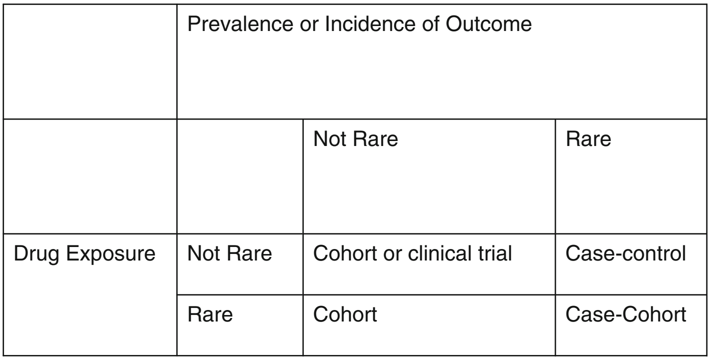
In addition, pharmacoepidemiology is growing around the world because of availability of electronic databases (e.g. claims, medical records), advances in computers with more powerful software and hardware, and improvements in methodological approaches to deal with various types of confounding particularly confounding by indication [23].

This chapter will review the factors involved in the selection of the type of pharmacoepidemiologic study design, and advantages and disadvantages of these designs. Since other chapters describe randomized clinical trials in detail, we will focus on observational studies.

Selection of Study Design

Many of the considerations necessary to determine the optimal study in PE are similar to those discussed in prior chapters; however, a brief review here will serve as a necessary reminder. Thus, before one can select the appropriate study design, one needs an appropriate research question that includes the objective and the purpose of the study (as is true for traditional epidemiologic studies). There is a consensus that an appropriate research question includes information about the exposure, outcome, and the population of interest, and they are included in the protocol. For example, an investigator might be interested in the question of whether there is an association of rosiglitazone with cardiac death in patients with type 2 diabetes mellitus. In this case, the exposure is the antidiabetic drug rosiglitazone, the outcome is cardiac death, and the population is a group of patients with type 2 diabetes. Although this may seem simplistic, it is surprising how many times it is unclear what the exact research question of a study is, and what the elements are which are under study.

The key elements for clearly stated objectives are keeping them SMART: Specific, Measurable, Appropriate, Realistic and Time-bound (SMART) [24]. An



	Prevalence or Incidence of Outcome		
		Not Rare	Rare
Drug Exposure	Not Rare	Cohort or clinical trial	Case-control
	Rare	Cohort	Case-Cohort

Fig. 12.1 Designs by frequency of exposure and outcome

objective is specific if it indicates the target; in other words, who and what is the focus of the research, and what outcomes are expected. By measurable, it is meant that the objective includes a quantitative measure. Appropriate, refers to an objective that is sensitive to target needs and societal norms, and realistic refers to an objective that includes a measure which can be reasonably achieved under the given conditions of the study. Finally, time-bound refers to an objective that clearly states the study duration. For example, a clearly stated objective might be: ‘to estimate the risk of rosiglitazone used as monotherapy on cardiac death in patients with type 2 diabetes treated between the years 2000–2007.’

In summary, in PE as in other areas of clinical research, clearly stated objectives are important in order to decide on the study design and analytic approach. That is, when a researcher has a clear idea about the research question and objective, it leads naturally to the optimal study design. Additionally, the investigator then takes into account the nature of the disease, the type of exposure, and available resources in order to complete the thought process involved in determining the optimal design and analysis approach. By the ‘nature of the disease’ it is meant that one is cognizant of the natural history of the disease from its inception to death. For example, a disease might be acute or chronic, and last from hours to years, and these considerations will determine whether the study needs to follow a cohort for weeks or for years in order to observe the outcome of interest. In PE research, the exposure usually refers to a drug or medication, and this could result in a study that could vary in duration (hours to years), frequency (constant or temporal) and strength (low vs. high dose). All of these aforementioned factors will have an impact on the selection of the design and the conduct of the study. In addition, a researcher might be interested in the effect of an exposure at one point in time (e.g. cross-sectional) vs. an exposure over long periods of time (e.g. cohort, case-control).

Since almost every research question can be approached using various designs, the investigator needs to consider both the strengths and weaknesses of each design in order to come to a final decision. For example, if an exposure is rare, the most efficient design is a cohort study (provided the outcome is common) but if the outcome is rare, the most efficient design is a case-control study (provided the exposure is common). If both the outcome and exposure are rare, a case-cohort design might be appropriate where odds ratio might be calculated with exposure data from a large reference cohort (Fig. 12.1).

Study Designs Common in PE

Table 12.1 demonstrates the study designs frequently used in PE research. Observational designs are particularly useful to study unintended drug effects in the postmarketing phase of the drug cycle. It is also important to consider the comparative effectiveness trial that is used in postmarketing research (see Chap. 5).

Effectiveness trials can be randomized or not randomized, and they are characterized by the head-to-head comparison of alternative treatments in large heterogeneous populations, imitating clinical practice [25–27]. As it is mentioned in Chap. 3, randomized clinical trials provide the most robust evidence, but they have often limited utility in daily practice because of selective population (e.g. specific disease severity, number of comorbidities and concomitant medications), small sample size, low drug doses, short follow-up period, and highly controlled environment [28].

Descriptive Observational Studies

Recall that these are predominantly hypothesis generating studies where investigators try to recognize or to characterize a problem in a population. In PE research, for example, investigators might be interested in recognizing unknown adverse effects, in knowing how a drug is used by specific populations, or how many people might be at risk of an adverse drug event. As a consequence, these studies do not generally

Table 12.1 Classification of postmarketing studies

I. Descriptive observational studies
A. Case report
B. Case series
C. Ecologic studies
D. Cross-sectional studies
II. Analytical studies
Observational studies
A. Case-control studies
B. Cross-sectional studies
C. Cohort studies
D. Hybrid studies
1. Nested case-control studies
2. Case-cohort studies
3. Case-crossover studies
4. Case-time studies
Interventional studies
A. Controlled clinical trials
B. Randomized, control clinical trials
C. N of trials
D. Simplified clinical trials
E. Community trial

measure associations; rather, they use measures of frequency such as proportions, rate, risk and prevalence.

Case Report

Case reports are descriptions of the history of a single patient who has been exposed to a medication and experiences a particular and unexpected effect, whether that effect is beneficial or harmful. In contrast to traditional research, in pharmacoepidemiologic research, case reports have a privileged place, because they can be the first signal of an adverse drug event, or the first indication for the use of a drug for conditions not previously approved (off-label indications by the regulatory agency e.g. Food and Drug Administration). As an example, case reports were used to communicate unintended adverse events such as phocomelia associated with the use of thalidomide [29]. Case reports also make up the key element for spontaneous reporting systems such as MedWatch, The FDA Safety Information and Adverse Event Reporting Program. The MedWatch program allows providers, consumers and manufacturers to report serious problems that they suspect are associated with the drugs and medical devices they prescribe, dispense, or use. By law, manufacturers, when they become aware of any adverse effect, must submit a case report form of serious unintended adverse events that have not been listed in the drug labeling within 15 calendar days [30].

Case Series

Case series is essentially a collection of ‘case reports’ that share some common characteristics such as being exposed to the same drug; and, in which same outcome is observed. Frequently, case series are part of phase IV postmarketing surveillance studies, and pharmaceutical companies may use them to obtain more information about the effect, beneficial or harmful, of a drug. For example, Humphries, et al. reported a case series of cimetidine carried out in its postmarketing phase, in order to determine if cimetidine was associated with agranulocytosis [31]. The authors followed new cimetidine users, and ultimately found no association with agranulocytosis. Often, case series characterize a certain drug-disease association in order to obtain more insight into the clinicopathological pattern of an adverse effect; such as, hepatitis occurring as a result of exposure to nitrofurantoin [32]. The main limitation of case series is that they do not include a comparison group(s). The lack of a comparison group is critical, and the result is that is difficult to determine if the drug effect is greater, the same or less than the expected effect in a specific population (a situation that obviously complicates the determination of causality).

Ecologic Studies

Ecologic studies evaluate secular trends and are studies where trends of drug-related outcomes are examined over time or across countries. In these studies, data from a single region can be analyzed to determine changes over time; or, data from a single time period can be analyzed to compare one region vs. another. Since ecologic studies do not provide data on individuals (rather they analyze data based on study groups), it is not only impossible to adjust for confounding variables; but, it does not reveal whether an individual with the disease of interest actually used the drug (this is termed the ecologic fallacy). In ecologic studies, sales, marketing, and claims databases are commonly used. For example, one study compared urban vs. the rural areas in Italy using drug sales data to assess for regional differences in the sales of tranquilizers [33, 34]. For the reasons given above, ecologic studies are limited in their ability to associate a specific drug with an outcome; and, invariably there are usually other factors that could also explain the outcome.

Cross-Sectional Studies

Cross-sectional studies are particularly useful in drug utilization studies and in prescribing studies, because they can present a picture of how a drug is actually used in a population or how providers are actually prescribing medications. Cross-sectional studies can be descriptive or analytical. Cross-sectional studies are considered descriptive in nature when they describe the ‘big’ picture about the use of a drug in a population, and the information about the exposure and the outcome are obtained at the same point in time. Cross sectional designs are used in drug utilization studies because these studies are focused on prescription, dispensing, administration of medication, marketing, and distribution; and, also address the use of drugs at a societal level, with special emphasis on the drugs resultant effect on medical, social, and economic consequences. Cross-sectional studies in PE are particularly important to determine how specific groups of patients, e.g. elderly, children, minorities, pregnant, etc. are using medications. As an example, Paulose-Ram et al. analyzed the U.S. National Health and Nutrition Examination Survey (NHANES) from 1988 to 1994 in order to estimate the frequency of analgesic use in a nationally representative sample from the U.S. From this study it was estimated that 147 million adults used analgesics monthly, women and Caucasians used more analgesics than men and other races, and more than 75 % of the use was over the counter [35].

Analytical Studies

Analytic studies, by definition, have a comparison group and as such are more able to assess an association or a relationship between an exposure and an outcome. If the investigator is able to allocate the exposure, the analytical study is considered to be an interventional study; while if the investigator does not allocate the exposure; the study is considered observational or non-experimental (or non-interventional). Analytical observational pharmacoepidemiologic studies quantify beneficial or adverse drug effects using measures of association such as rate, risk, odds ratios, rate ratios, or risk difference. Analytic pharmacoepidemiologic studies are particularly important when there are uncommon or delayed adverse events because clinical trials would be impractical and/or unfeasible especially if event rates are lower than 1:2,000 or 1:3000 [36].

Cross-Sectional Studies

Cross-sectional studies can be analytical if they are attempting to demonstrate an association between an exposure and an outcome. For example, Paulose-Ram et al. used the NHANES III data to estimate the frequency of psychotropic medication used among Americans between 1988 and 1994; and, to estimate if there was an association of sociodemographic characteristics with psychotropic medication use. They found that psychotropic medications were associated with low socioeconomic status, lack of high school education, and whether subjects were insured [37]. The problem with analytical cross-sectional studies is that it is often unknown whether the exposure really precedes the outcome because both are measured at the same point in time. This is obviously important since if the exposure does not precede the outcome, it can not be the cause of that outcome. This is especially important in cases of chronic disease where it may be difficult to ascertain which drugs preceded the onset of that disease.

Case-Control Studies (or Case-Referent Studies)

Case control and cohort studies are designs where participants are selected based on the outcome (case-control) or on the exposure (cohort) Fig. 12.2. In PE case-control studies, the odds of drug use among cases (the ratio exposed cases/unexposed cases) are compared to the odds of drug use among non cases (the ratio exposed controls/unexposed controls). The case-control design is particularly desirable when one wants to study multiple determinants of a single outcome [38]. The case-control design is a particularly efficient study when the outcomes are rare, since the design guarantees a sufficient number of cases. For example, Ibanez et al. designed a case-control study to estimate the association of non-steroidal anti-inflammatory drugs (NSAID) (common exposure) with end-stage renal disease (a rare outcome). In this study, the cases were patients entering a local dialysis program from 1995 to 1997 as a result of end-stage renal disease; while controls, were selected from the hospital where the case was first diagnosed (in addition, the controls did not have conditions associated with NSAID use). Information on previous use of NSAID drugs (exposure) was then obtained in face-to-face interviews (which, by the way, might introduce bias – this type of bias may be prevented if prospectively gathered prescription data are available, although for NSAIDs the over-the-counter use is almost never registered on an individual basis).

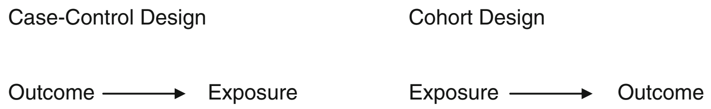


Fig. 12.2 Case-control and cohort designs

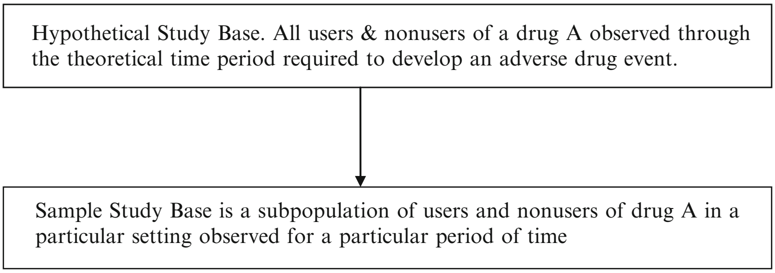

Hypothetical Study Base. All users & nonusers of a drug A observed through the theoretical time period required to develop an adverse drug event.
Sample Study Base is a subpopulation of users and nonusers of drug A in a particular setting observed for a particular period of time

Fig. 12.3 Study base and sample study base

As implied above, case-control studies are vulnerable to selection, information and confounding bias. For example, selection bias can occur when the cases enrolled in the study have a drug use profile that is not representative of all cases. For instance, selection bias occurs if cases are identified from hospital data and if people with the medical condition of interest are more likely to be hospitalized if they used the drug (than if they did not). Selection bias may also occur by selective nonparticipation in the study, or when controls enrolled in a study have a drug use profile that differs from that of the 'sample study base' (Fig. 12.3). Selection bias can then be minimized if controls are selected from the same source population (study base) as the cases [39, 40].

Since the exposure information in case-control studies is frequently obtained retrospectively-through medical records, interviews, and self-administered questionnaires, case-control studies are often subject to information bias. Most information bias pertains to recall and measurement bias. Recall bias may occur, for example, when interviewed cases remember more details about drug use than non-cases. The use of electronic pharmacy databases, with complete information about drug exposure, could reduce this type of bias. Finally, an example of measurement or diagnostic bias occurs when researchers partly base the diagnosis of interpretation of the diagnosis on knowledge of the exposure status of the study subjects.

Cohort Studies

Recall, that in cohort studies, participants are recruited based on the exposure and they are followed up over time while studying differences in their outcome. In PE cohort studies, users of a drug are compared to nonusers or users of other drugs with

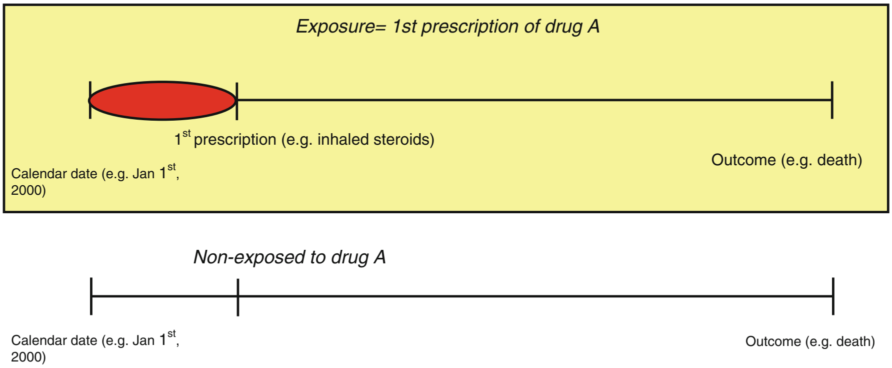

Exposure= 1st prescription of drug A
1st prescription (e.g. inhaled steroids)
Calendar date (e.g. Jan 1st, 2000)
Outcome (e.g. death)
Non-exposed to drug A
Calendar date (e.g. Jan 1st, 2000)
Outcome (e.g. death)

Fig. 12.4 Immortal time bias in exposed (Study) and non-exposed (Control) groups (Adapted from Refs. [44–49])

respect to rate or risk of an outcome. PE cohort studies are particularly efficient for rarely used drugs, or when there are multiple outcomes from a single exposure. The cohort study design then allows for establishing a temporal relationship between the exposure and the outcome because drug use precedes the onset of the outcome. In cohort studies, selection bias is generally less likely to occur than in case-control designs. Selection bias is less likely to occur, for example, when the drug use profile of the sample study base is similar to that of subjects enrolled in the study.

The disadvantages of cohort studies include the need for large number of subjects (unless the outcome is common, cohort studies are potentially uninformative for rare outcomes – especially those which require a long observation period); they are generally more expensive than other designs, particularly if active data collection is needed. In addition, they are vulnerable to bias if a high number of participants are lost during the follow-up (high drop-out rate). Finally, for some retrospective cohort studies, information about confounding factors might be limited or unavailable. With retrospective cohort studies, for example, the study population is frequently dynamic because the amount of time during which a subject is observed varies from subject to subject. PE retrospective cohort studies are frequently performed with information from automated databases with reimbursement or health care information (e.g. Veterans Administration database, Saskatchewan database, PHARMO database).

A special bias exists with cohort studies, the immortal time bias, which can occur when, as a result of the exposure definition, a subject, cannot incur the outcome event of interest during the follow up. For example, if an exposure is defined as the first prescription of drug 'A', and the outcome is death, the period of time from the calendar date to the first prescription where the outcome does not occur is the immortal time bias (red oval in Fig. 12.4). If during that period, the outcome occurs (e.g. death), then the subject won't be classified as part of the study group, rather, that subject will be part of the control group. This type of bias was described in the seventies when investigators compared the survival time of individuals receiving a heart transplant (study group) vs. those who were candidates but did not receive the transplant (control group). They found longer survival in the study group [41, 42]. A reanalysis of data demonstrated that there was a waiting time from diagnosis of cardiac disease to the heart transplant, where patients were 'immortal' because if they died before the heart transplant, they were part of the control group [43]. This concept was adopted in pharmacoepidemiology research and since then, many published studies have been described with this type of bias [44–49]. (Fig. 12.4).

As prior mentioned, the consequence of this immortal time bias is the spurious appearance of a better outcome in the study group such as lower death rates. In other words, there is an underestimation of person-time without a drug treatment leading to an overestimation of a treatment effect [50]. One of the techniques to avoid immortal time bias is time-dependent drug exposure analysis [51].

Hybrid Studies

In PE research, hybrid designs are commonly used to study drug effects and drug safety. These designs combine several standard epidemiologic designs with resulting increased efficiency. In these studies, cases are selected on the basis of the outcome; and, drug use is compared with the drug use of several different types of comparison groups (see Table 12.2). These designs include: nested-case control studies, case-cohort design, case-crossover design, case-time-control design, and self-controlled case series [52].

Nested Case-Control Studies

Recall that a nested case-control study refers to a case-control study which is nested in a cohort study or RCT. In PE, nested case-control studies, a defined population is followed for a period of time until a number of incident cases of a

Table 12.2 A description of some hybrid postmarketing study designs

Design	Control group
Nested case-control	Subjects in the same cohort, without the case condition
Case-cohort	A sample of the cohort at baseline (may include later cases)
Case-crossover	Cases, at an earlier time period
Case-time-control	Cases, at an earlier time period but time effect is considered
Self-controlled case series	Cases are their own controls

disease or an adverse drug reaction is identified. If the case-control study is nested in a cohort with prospectively gathered data on drug use, recall bias is no longer a problem. In PE as in other clinical research, nested case-control studies are used when the outcome is rare or the outcome has long induction time and latency. Frequently, this type of design is used when there is the need to use stored biological samples and additional information on drug use and confounders are needed. When it is inefficient to collect the aforementioned data for the complete cohort, (a common occurrence) a nested case-control study is desirable.

Case-Cohort Studies

Recall that this type of study is similar to a nested case-control design, except the exposure and covariate information is collected from all cases, whereas controls are a random representative sample selected from the original cohort [53, 54]. Case-cohort studies are recommended in the presence of rare outcomes or when the outcome has a long induction time and latency, but especially when the exposure is rare (if the exposure in controls is common, a case-control study is preferable). In PE case-cohort studies, the proportion of drug use in cases is compared to the proportion of drug use in the reference cohort (which may include cases). An example of the use of this design was to evaluate the association between immunosuppressive therapy (cyclophosphamide, azathioprine and methotrexate) and haematological changes in lung cancer, in patients with systemic lupus erythematosus (this was based on a lupus erythematosus cohort from centers in North America, Europe and Asia, where exposure and covariate information for all cases was collected). Cases were defined as SLE, with invasive cancers discovered at each center after entry into the lupus cohort; and, the index time for each risk set was the date of the case's cancer occurrence. Controls were obtained from a random sample of the cohort (10 % of the full cohort) and they represented cancer free patients up to the index time. Authors found that immunosuppressive therapy may contribute to an increased risk of hematological malignancies [55].

Case-Crossover Studies

Recall that the case-crossover design was proposed by Maclure, and in this design only cases that have experienced an outcome are considered. In that way, each case contributes one case window and one or more control windows at various time periods, and for the same patient. In other words, control subjects are the same as cases, just at an earlier time, so cases serve as own controls (see Chap. 4) [56, 57]. This type of

Exposed and Unexposed Periods in the Same Subject

Unexposed time period	Exposed time period	Unexposed time period	Exposed time period	Unexposed time period
Control time1	Case	Control time2	Case	Control time3

Fig. 12.5 Case-crossover design

design is particularly useful when a disease does not vary over time and when exposures are transient, brief and acute [56, 58]. The case-crossover design contributes to the elimination of control selection bias and avoids difficulties in selecting and enrolling controls. However, case crossover designs are not suitable for studying chronic conditions [59]. In PE, case-crossover studies might compare the odds of drug use at a time close to onset of a medical condition compared with odds at an earlier time (Fig. 12.5).

Case-crossover designs have been used to assess the acute risks of vehicular accidents associated with the use of benzodiazepines [60] and also to study changes in medication use associated with epilepsy-related hospitalization. In this latter study, Handoko et al. used the PHARMO database from 1998 to 2002. For each patient, changes in medication in a 28-day window before hospitalization, were compared with changes in four earlier 28-day windows; and, pattern of drug use, dosages, and interaction with medications were analyzed. Investigators found that patients starting with three or more new non antiepileptic drugs had a five times higher risk of epilepsy-related hospitalization [61]. In case-crossover designs, conditional logistic regression analysis is classically used to assess the association between event and exposure [62, 63].

Case-Time-Control Studies

The case-time control design was proposed by Suissa [64] to control for confounding by indication. In this design subjects from a conventional case-control design are used as their own controls. This design is an extension of the case-crossover design but it takes into account the time effect, particularly the variation in the drug use over time. This type of design is recommended when an exposure varies over time and when there are two or more points measured at different times, and it is expected to be able to separate the drug effect from the disease severity. Something to consider is that the same precautions used in case-crossover designs should also be taken into account in case-time-control designs, and the exposures of control subjects must be measured at the same points in calendar time as their cases.

Self-Controlled Case Series

In this design, case series are used to study the temporal association between a time-varying exposure and an adverse event (acute event) using data on cases only [52]. In this case, the effect of exposure is transitory and limited to a certain risk period, and then it returns to baseline. For example, if there is interest in studying thrombocytopenia associated with a vaccine administered at specific age, the risk period is limited to that age period. The assumptions of self-controlled case series include: the occurrence of an event must not alter the probability of subsequent exposure, the occurrence of the event must not affect the observation period and recurrent events should be independent or if they are not but the event is rare, only the first event can be used. The advantage of this design is that cases are their own controls which imply an adjustment of confounders (e.g. socioeconomic factors). In addition, it reduces the effort and cost of data collection [52].

Biases in PE

In PE, a special type of bias (confounding by indication) occurs when those subjects who receive the drug have an inherently different prognosis from those who do not receive the drug. If the indication for treatment is an independent risk factor for the study outcome, the association of this indication with the prescribed drug may cause confounding by indication. A variant of confounding by indication (confounding by severity) may occur if a drug is prescribed selectively to patients with specific disease severity profiles [65]. Some hybrid designs and statistical techniques have been proposed to control for confounding by indication. In terms of statistical techniques, it has been proposed that one use multivariable model risk adjustment, propensity score risk adjustment, propensity-based matching and instrumental variable analysis to control for confounding by indication. Multivariable model risk adjustment is a conventional modeling approach that incorporates all known confounders into the model. Controlling for those covariates produces a risk-adjusted treatment effect and removes overt bias due to those factors [66].

Propensity score risk adjustment is a technique used to adjust for nonrandom treatment assignment. It is a conditional probability of assignment to a particular treatment given a set of observed patient-level characteristics [67, 68]. In this technique, a score is developed for each subject based on a prediction equation and the subject's value of each variable is included in the prediction equation [69], and it is a scalar summary of all observed confounders. Within propensity score strata, covariates in treated and non-treated groups are similarly distributed, so the stratification using propensity score strata is claimed to remove more than 90 % of the overt bias due to the covariates used to estimate the score [70, 71]. Unknown biases can be partially removed only if they are correlated with covariates already measured and included in the model to compute the score [72–74].

Instrumental variable analysis is an econometric method used to remove the effects of hidden bias in observational studies [75, 76]. Instrumental variables are highly correlated with treatment and they do not independently affect the outcome. Therefore, they are not associated with patient health status. Instrumental variable analysis compared groups of patients that differ in likelihood of receiving a drug [77].

Summary

In pharmacoepidemiology research as is true for traditional research, the selection of an appropriate study design requires the consideration of various factors such as the frequency of the exposure and outcome, and the population under study. Investigators frequently need to weigh the choice of a study design with the quality of information collected along with its associated costs. In fact, new pharmacoepidemiologic designs are being developed to improve study efficiency.

Pharmacoepidemiology is not a new discipline, but it is currently recognized as one of the most challenging and growing areas in research, and many techniques and methods are being tested to confront those challenges. Pharmacovigilance (See Chap. 5) as a part of pharmacoepidemiology is of great interest for decision makers, researchers, providers, manufacturers and the public, because of concerns about drug safety. Therefore, we should expect in the future, the development of new methods to assess the risk/benefit ratios of medications.

References

1. Strom B, Kimmel S. Textbook of pharmacoepidemiology. Hoboken: Wiley; 2006.
2. Miller JL. Troglitazone withdrawn from market. Am J Health Syst Pharm. 2000;57:834.
3. Gale EA. Lessons from the glitazones: a story of drug development. Lancet. 2001;357:1870–5.
4. Scheen AJ. Thiazolidinediones and liver toxicity. Diabetes Metab. 2001;27:305–13.
5. Glessner MR, Heller DA. Changes in related drug class utilization after market withdrawal of cisapride. Am J Manag Care. 2002;8:243–50.
6. Griffin JP. Prepulsid withdrawn from UK & US markets. Adverse Drug React Toxicol Rev. 2000;19:177.
7. Graham DJ, Staffa JA, Shatin D, Andrade SE, Schech SD, La Grenade L, et al. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. JAMA. 2004;292:2585–90.
8. Piorkowski Jr JD. Bayer's response to "potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis". JAMA. 2004;292:2655–7. discussion 2658–9.
9. Strom BL. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: a counterpoint. JAMA. 2004;292:2643–6.
10. Wooltorton E. Bayer pulls cerivastatin (Baycol) from market. Can Med Assoc J. 2001;165:632.
11. Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet. 2004;364:2021–9.

12. Sibbald B. Rofecoxib (Vioxx) voluntarily withdrawn from market. Can Med Assoc J. 2004;171:1027–8.
13. Wong M, Chowienczyk P, Kirkham B. Cardiovascular issues of COX-2 inhibitors and NSAIDs. Aust Fam Physician. 2005;34:945–8.
14. Antoniou K, Malamas M, Drosos AA. Clinical pharmacology of celecoxib, a COX-2 selective inhibitor. Expert Opin Pharmacother. 2007;8:1719–32.
15. Sun SX, Lee KY, Bertram CT, Goldstein JL. Withdrawal of COX-2 selective inhibitors rofecoxib and valdecoxib: impact on NSAID and gastroprotective drug prescribing and utilization. Curr Med Res Opin. 2007;23:1859–66.
16. Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. Am J Epidemiol. 2005;162:404–14.
17. Dubach UC, Rosner B, Sturmer T. An epidemiologic study of abuse of analgesic drugs. Effects of phenacetin and salicylate on mortality and cardiovascular morbidity (1968 to 1987). N Engl J Med. 1991;324:155–60.
18. Elseviers MM, De Broe ME. A long-term prospective controlled study of analgesic abuse in Belgium. Kidney Int. 1995;48:1912–9.
19. Morlans M, Laporte JR, Vidal X, Cabeza D, Stolley PD. End-stage renal disease and non-narcotic analgesics: a case-control study. Br J Clin Pharmacol. 1990;30:717–23. PMC1368172.
20. Murray TG, Stolley PD, Anthony JC, Schinnar R, Hepler-Smith E, Jeffreys JL. Epidemiologic study of regular analgesic use and end-stage renal disease. Arch Intern Med. 1983;143:1687–93.
21. Perneger TV, Whelton PK, Klag MJ. Risk of kidney failure associated with the use of acetaminophen, aspirin, and nonsteroidal antiinflammatory drugs. N Engl J Med. 1994;331:1675–9.
22. International Society of Pharmacoepidemiology (ISPE). Guidelines for Good Pharmacoepidemiology Practices (GPP). Pharmacoepidemiol Drug Saf. 2008;17:200–8.
23. Avorn J. The promise of pharmacoepidemiology in helping clinicians assess drug risk. Circulation. 2013;128:745–8. doi:10.1161/CIRCULATIONAHA.113.003419.
24. Piotrow PT, Kincaid DL, Rani M, Lewis G. Communication for social change. Baltimore: The Rockefeller Foundation/Johns Hopkins Center for Communication Programs; 2002.
25. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). JAMA. 2002;288:2981–97.
26. Pilote L, Abrahamowicz M, Rodrigues E, Eisenberg MJ, Rahme E. Mortality rates in elderly patients who take different angiotensin-converting enzyme inhibitors after acute myocardial infarction: a class effect? Ann Intern Med. 2004;141:102–12.
27. Schneider LS, Tariot PN, Dagerman KS, Davis SM, Hsiao JK, Ismail MS, et al. Effectiveness of atypical antipsychotic drugs in patients with Alzheimer's disease. N Engl J Med. 2006;355:1525–38.
28. Schneeweiss S. Developments in post-marketing comparative effectiveness research. Clin Pharmacol Ther. 2007;82:143–56. PMC2905665.
29. Mellin GW, Katzenstein M. The saga of thalidomide. Neuropathy to embryopathy, with case reports of congenital anomalies. N Engl J Med. 1962;267:1238–44.
30. Food and Drug Administration. Medwatch Website. www.fda/gov/medwatch. Accessed 20 Aug 2007.
31. Humphries TJ, Myerson RM, Gifford LM, Aeugle ME, Josie ME, Wood SL, et al. A unique postmarket outpatient surveillance program of cimetidine: report on phase II and final summary. Am J Gastroenterol. 1984;79:593–6.
32. Stricker BH, Blok AP, Claas FH, Van Parys GE, Desmet VJ. Hepatic injury associated with the use of nitrofurans: a clinicopathological study of 52 reported cases. Hepatology. 1988;8:599–606.

33. Martin A, Leslie D. Trends in psychotropic medication costs for children and adolescents, 1997–2000. Arch Pediatr Adolesc Med. 2003;157:997–1004.
34. Williams P, Bellantuono C, Fiorio R, Tansella M. Psychotropic drug use in Italy: national trends and regional differences. Psychol Med. 1986;16:841–50.
35. Paulose-Ram R, Hirsch R, Dillon C, Losonczy K, Cooper M, Ostchega Y. Prescription and non-prescription analgesic use among the US adult population: results from the third National Health and Nutrition Examination Survey (NHANES III). Pharmacoepidemiol Drug Saf. 2003;12:315–26.
36. U.S. Food and Drug Administration. Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment. March 2005.
37. Paulose-Ram R, Jonas BS, Orwig D, Safran MA. Prescription psychotropic medication use among the U.S. adult population: results from the third National Health and Nutrition Examination Survey, 1988–1994. J Clin Epidemiol. 2004;57:309–17.
38. Strom B. Study designs available for pharmacoepidemiology studies. In: Pharmacoepidemiology. 3rd ed. Wiley; 2000.
39. International Agranulocytosis and Aplastic Anemia Study Group. Risks of agranulocytosis and aplastic anemia: a first report of their relation to drug use with special reference to analgesics. JAMA. 1986;256:1749–57.
40. Wilcox AJ, Baird DD, Weinberg CR, Hornsby PP, Herbst AL. Fertility in men exposed prenatally to diethylstilbestrol. N Engl J Med. 1995;332:1411–6.
41. Clark DA, Stinson EB, Griepp RB, Schroeder JS, Shumway NE, Harrison DC. Cardiac transplantation in man. VI. Prognosis of patients selected for cardiac transplantation. Ann Intern Med. 1971;75:15–21.
42. Messmer BJ, Nora JJ, Leachman RD, Cooley DA. Survival-times after cardiac allografts. Lancet. 1969;1:954–6.
43. Gail MH. Does cardiac transplantation prolong life? A reassessment. Ann Intern Med. 1972;76:815–7.
44. Donahue JG, Weiss ST, Livingston JM, Goetsch MA, Greineder DK, Platt R. Inhaled steroids and the risk of hospitalization for asthma. JAMA. 1997;277:887–91.
45. Fan VS, Bryson CL, Curtis JR, Fihn DS, Bridevaux PO, McDonell MD, et al. Inhaled corticosteroids in chronic obstructive pulmonary disease and risk of death and hospitalization: time-dependent analysis. Am J Respir Crit Care Med. 2003;168:1488–94.
46. Kiri VA, Vestbo J, Pride NB, Soriano JB. Inhaled steroids and mortality in COPD: bias from unaccounted immortal time. Eur Respir J. 2004;24:190–1; author reply 191–2.
47. Mamdani M, Rochon P, Juurlink DN, Anderson GM, Kopp A, Naglie G, et al. Effect of selective cyclooxygenase 2 inhibitors and naproxen on short-term risk of acute myocardial infarction in the elderly. Arch Intern Med. 2003;163:481–6.
48. Suissa S. Observational studies of inhaled corticosteroids in chronic obstructive pulmonary disease: misconstrued immortal time bias. Am J Respir Crit Care Med. 2006;173:464; author reply 464–5.
49. Suissa S. Immortal time bias in observational studies of drug effects. Pharmacoepidemiol Drug Saf. 2007;16:241–9.
50. Suissa S. Effectiveness of inhaled corticosteroids in chronic obstructive pulmonary disease: immortal time bias in observational studies. Am J Respir Crit Care Med. 2003;168:49–53.
51. Time-varying explanatory variables. In: Clayton D, Hills M, editors. Statistical models in epidemiology. Oxford: Oxford University Press; 1993. p. 307–18.
52. Whitaker HJ, Hocine MN, Farrington CP. The methodology of self-controlled case series studies. Stat Methods Med Res. 2009;18:7–26. doi:10.1177/0962280208092342.
53. Sato T. Risk ratio estimation in case-cohort studies. Environ Health Perspect. 1994;102:53–6. PMC1566546.
54. van der Klauw MM, Stricker BH, Herings RM, Cost WS, Valkenburg HA, Wilson JH. A population based case-cohort study of drug-induced anaphylaxis. Br J Clin Pharmacol. 1993;35:400–8. PMC1381551.

55. Bernatsky S, Boivin JF, Joseph L, Gordon C, Urowitz M, Gladman D, et al. The relationship between cancer and medication exposures in systemic lupus erythematosus: a case-cohort study. Ann Rheum Dis. 2008;67:74–9.
56. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol. 1991;133:144–53.
57. Maclure M, Mittleman MA. Should we use a case-crossover design? Annu Rev Public Health. 2000;21:193–221.
58. Marshall RJ, Jackson RT. Analysis of case-crossover designs. Stat Med. 1993;12:2333–41.
59. Donnan PT, Wang J. The case-crossover and case-time-control designs in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2001;10:259–62.
60. Barbone F, McMahon AD, Davey PG, Morris AD, Reid IC, McDevitt DG, et al. Association of road-traffic accidents with benzodiazepine use. Lancet. 1998;352:1331–6.
61. Handoko KB, Zwart-van Rijkom JE, Hermens WA, Souverein PC, Egberts TC. Changes in medication associated with epilepsy-related hospitalisation: a case-crossover study. Pharmacoepidemiol Drug Saf. 2007;16:189–96.
62. Greenland S. A unified approach to the analysis of case-distribution (case-only) studies. Stat Med. 1999;18:1–15.
63. Scneeweiss S, Stürmer T, Maclure M. Case-crossover and case = time-control designs as alternatives in pharmacoepidemiologic research. Pharmacoepidemiol Drug Saf. 1997;6:S51–9.
64. Suissa S. The case-time-control design. Epidemiology. 1995;6:248–53.
65. Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. Am J Epidemiol. 1999;149:981–3.
66. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA. 2007;297:278–85.
67. D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17:2265–81.
68. Morant SV, Pettitt D, MacDonald TM, Burke TA, Goldstein JL. Application of a propensity score to adjust for channeling bias with NSAIDs. Pharmacoepidemiol Drug Saf. 2004;13: 345–53.
69. Ahmed A, Husain A, Love TE, Gambassi G, Dell'Italia LJ, Francis GS, et al. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. Eur Heart J. 2006;27:1431–9. PMC2443408.
70. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
71. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc. 1984;79:516–24.
72. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. Stat Med. 2005;24:1563–78.
73. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Intern Med. 2002;137:693–5.
74. Harrell FE. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Springer; 2001.
75. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. JAMA. 1994;272:859–66.
76. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. Annu Rev Public Health. 1998;19:17–34.
77. Harris KM, Remler DK. Who is the marginal patient? Understanding instrumental variables estimates of treatment effects. Health Serv Res. 1998;33:1337–60. PMC1070319.

Chapter 13
Implementation Research: Beyond the Traditional Randomized Controlled Trial

Amanda S. Mixon, Lauren Powell, and Carlos A. Estrada

> *I think when people look back at our time, they will be amazed at one thing more than any other. It is this-that we do know more about ourselves now than people did in the past, but that very little if this knowledge has been put into effect.*
>
> D Lessing
> http://daronlarson.blogspot.com/2006/12/prisons-we-choose-to-live-inside.html

Abstract Implementation research is a new scientific discipline emerging from the recognition that the public does not derive sufficient or rapid benefit from advances in the health sciences (Berwick DM, JAMA 289:1969–1975, 2003; Lenfant C, N Engl J Med 349:868–874, 2003). One often-quoted estimate claims that it takes an average of 17 years for even well-established clinical knowledge to be fully adopted into routine practice (Kiefe CI, Sales A, J Gen Intern Med 21(Suppl 2):S67–S70, 2006). In this chapter, we will discuss particular barriers to evidence implementation, present tools for implementation research, and provide a framework for designing implementation

A.S. Mixon, M.D., M.S., MSPH (✉)
Geriatric Research Education and Clinical Center (GRECC); Section of Hospital Medicine, Department of Veterans Affairs, Tennessee Valley Healthcare System, Vanderbilt University, Nashville, TN, USA
e-mail: Amanda.S.Mixon@Vanderbilt.Edu

L. Powell, BS
Department of Quantitative Health Sciences, Graduate Research Assistant, Division of Preventive and Behavioral Medicine, University of Massachusetts Medical School, Worcester, MA, USA

C.A. Estrada, M.D., M.S.
General Internal Medicine, University of Alabama at Birmingham, Birmingham, AL, USA

Veterans Affairs Quality Scholars Program, Birmingham VA Medical Center, Birmingham, AL, USA

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_13,

research studies, emphasizing the randomized trial. The reader is advised that this chapter only provides a basic introduction to several concepts for which new approaches are rapidly emerging. Therefore, our goal is to stimulate interest and promote additional in-depth learning for those who wish to develop new implementation research projects or better understand this exciting field.

Keywords Implementation research tools • Rogers Diffusion Theory • Translational barriers • Academic detailing • Pay-for-performance

Introduction

Overview and Definition of Implementation Research

Implementation research is an emergent discipline born from the recognition that the public does not derive sufficient or rapid benefit from advances in the health sciences [1, 2]. Implementation research bridges the gap between scientific knowledge and its application to daily practice with the overall purpose of improving the health of individuals and populations. One often-quoted estimate claims that it takes an average of 17 years for even well-established clinical knowledge to be fully adopted into routine practice [3]. In addition, approximately half of trials funded by the National Institutes of Health were published in peer-reviewed publications two and a half years after study completion [4].

For example, in 2000, only one-third of patients with coronary artery disease received aspirin when no contraindications to its use were present; [2] furthermore, a landmark study estimated that the American public was only receiving about 55 % of recommended care [5]. Implementation research definitions are shown in Table 13.1.

A glossary of terms used in implementation research is now available [13, 14]. The definition of implementation research may be expanded to encompass work that promotes patient safety and eliminates racial and ethnic disparities in health care. Health disparities implementation research aims to identify strategies to close gaps in health care through culturally-appropriate interventions for patients, clinicians, health care systems, and populations [15–18]. Under-represented populations make up a significant portion of the U.S. population, shoulder a disproportionate burden of disease, and receive inadequate care [19]. According to the U.S. National Institute of Health (NIH), '*dissemination and implementation research intends to bridge the gap between public health, clinical research, and everyday practice by building a knowledge base about how health information, interventions, and new clinical practices and policies are transmitted and translated for public health and health care service use in specific settings*' [6].

Gaps in health care may be classified as 'errors of omission,' (failure to provide necessary care [20]) and 'errors of commission,' such as the delivery of unnecessary or inappropriate care which causes harm. A landmark report from the Institute of

Table 13.1 Implementation research – definitions and terms

‘is the scientific study of methods to promote the integration [and rapid uptake] of research findings and evidence-based interventions into healthcare practice and policy, [and hence improve the health of individuals and populations]’ [6, 7]

‘…scientific investigations that support movement of evidence-based, effective health care approaches from the clinical knowledge base into routine use. …. Implementation science consists of a body of knowledge on methods to promote the systematic uptake of new or underused scientific findings into the usual activities…’ [8]

‘[Knowledge translation] is a dynamic and iterative process that includes synthesis, dissemination, exchange and ethically-sound application of knowledge to improve the health of Canadians, provide more effective health services and products and strengthen the health care system’ [9, 10]

‘is the scientific study of methods to promote the systematic uptake of clinical research findings and other evidence-based practices into routine practice, and hence to improve the quality (effectiveness, reliability, safety, appropriateness, equity, efficiency) of health care. It includes the study of influences on healthcare professional and organizational behavior’ [11]

‘is the systematic study of how a specific set of activities and designed strategies are used to successfully integrate an evidence-based public health intervention within specific settings (e.g., primary care clinic, community center, school)’ [12]

Medicine drew attention to patient safety and the concept of preventable injury [21]. Studies of patient safety have focused on ‘*medical error resulting in an inappropriate increased risk of iatrogenic adverse event(s) from receiving too much or hazardous treatment (overuse or misuse)’* [20].

For example, inappropriate antibiotic use may promote microbial resistance and cause unnecessary adverse events. Since 1999, public efforts have been underway to promote appropriate prescribing of antibiotics for acute respiratory infections (ARIs) [22]. Based on well-designed studies demonstrating no benefit, guidelines have long recommended against antibiotic use for acute bronchitis; [23, 24] however, physicians continue to prescribe antibiotics for patients diagnosed with ARIs. Although overall antibiotic use for ARIs declined between 1995 and 2002, use of broad-spectrum antibiotic prescriptions for ARIs increased [25]. A more recent implementation research project successfully used a multidimensional intervention in emergency departments to decrease antibiotic prescribing [26].

In response to what may be perceived as overwhelming evidence that thousands of lives are lost each year from errors of omission and commission, there have been strong national calls for health systems, hospitals, and physicians to adopt new approaches for moving evidence into practice, but rigorous supporting evidence is often lacking [27, 28].

As our understanding of implementation science is evolving, local clinicians and health systems must strive to improve the quality of care for every patient. Certain local decisions must be based on combinations of incomplete empiric evidence and personal experience. As with the clinician caring for the individual patient, every decision about local implementation cannot be guided by data from a randomized trial [29, 30]. However, a stronger evidence base is needed to inform wide-spread implementation efforts. Widespread implementation beyond evidence raises concern about unintended consequences and opportunity costs from public resources wrongly expended on ineffective interventions [30].

Implementation researchers use a variety of techniques, ranging from qualitative exploration to the controlled, group-randomized trial. For example, methods used in social, cognitive, and organizational psychology are also applicable to implementation research [31]. Berwick reminds us of the importance of understanding the mechanism and context through which implementation techniques exert their potential effects within complex human systems [32]. Berwick cautioned that important lessons may be lost through aggregation and rigorous scientific experimentation, challenging the implementation research community to reconsider the basic concept of evidence, itself.

Interventions for translating evidence into practice must operate in complex, poorly understood environments with multiple interacting components that may not be easily reducible to a clean, scientific formula. Therefore, we later present situational analysis as a framing device for implementation research. Nonetheless, in keeping with the theme of this book, we mainly focus on the randomized trial as one of the many critical tools for implementation research.

In summary, implementation research is an emerging body of scientific work seeking to close the gap between knowledge generated from the health sciences and routine practice, ultimately improving patient and population health outcomes. Implementation research, which encompasses the patient, clinician, health system, and community, may promote the use of needed services or the avoidance of unneeded services. Implementation research often focuses on patients who are vulnerable because of race/ethnicity or socioeconomic position. By its very nature implementation research is inter-disciplinary.

In this chapter, we discuss barriers to evidence implementation, present tools for implementation research, and provide a framework for designing implementation research studies. The reader is advised that this chapter only provides a basic introduction to several concepts for which new approaches are rapidly emerging. Therefore, our goal is to stimulate interest and promote additional in-depth learning for those who wish to develop new implementation research projects or better understand this exciting field.

Overcoming Barriers to Evidence Implementation

Successful implementation of evidence-based interventions largely depends on their fit with the preferences and priorities of those who shape, deliver, and participate in healthcare [33]. Although the conceptual basis for moving evidence into practice has not been fully developed, a solid grounding in relevant theory may be useful to those designing new implementation research projects [34]. Many conceptual models have been developed in other settings and subsequently adapted for translating evidence into practice [35]. For example, implementation researchers frequently apply Roger’s theory describing innovation diffusion. Rogers proposed three clusters of influence on the rapidity of innovation uptake is influenced by:

- Perceived advantages of the innovation
- The classification of new technology users according to rapidity of uptake; and
- Contextual factors [36].

First, potential users are unlikely to adopt an innovation that is perceived to be complex and inconsistent with their needs and cultural norms. Second, rapidity of innovation uptake often follows a sigmoid-shaped curve, with an initial period of slow uptake led by the ‘innovators.’ Next follows a more rapid period of uptake led by the early adopters, or ‘opinion leaders.’ During the last adoption phase, the rate of diffusion again slows as the few remaining ‘laggards’ or traditionalists adopt the innovation. Finally, contextual or environmental factors such as organizational culture exert a profound impact on innovation adoption, a concept that is explored in more detail in the following sections of this chapter.

Consistent with the model proposed by Rogers, multiple barriers often work synergistically to hinder the translation of evidence into practice [37]. Interventions often require significant time, money, and staffing. Implementation sites may experience difficulties in implementation as a result of limited resources, competing demands, and entrenched practices. For example, the intervention may have been developed and tested under circumstances that differ from those at the planned implementation site. Further, the implementation team may not adequately understand the environmental characteristics proposed by Roger’s diffusion theory as critical to the adoption of innovation. Because of such concerns a thorough environmental analysis is needed prior to widespread implementation efforts [37].

Building upon models proposed by Sung et al. [38] and Rubenstein et al. [8], Figure 13.1 depicts the translational barriers implementation research seeks to overcome. The first translational roadblock lies between basic science knowledge and clinical trial design. The second roadblock involves translation of knowledge gained from clinical trials into meaningful clinical guidance, which often takes the form of evidence-based guidelines.

The third roadblock specific to implementation science occurs between current clinical knowledge and routine practice, carrying important implications for individual practitioners, health care systems, communities, and populations. Given the expansive nature of this third roadblock, a multifaceted armamentarium of tools is required. One tool, industrial-style quality improvement, described below in more detail, operates at the level of the clinical microsystem, the smallest, front-line functional unit that actually delivers care to a patient [39]. Clinical microsystems consist of complex adaptive relationships among patients, providers, support staff, technology, and processes of care. To achieve sustainable success, researchers seeking to overcome this third translational barrier need to be effective advocates for changes in larger macrocosms of the healthcare system including local and governmental health policy. Finally, implementation research may inform clinical trials and basic science.

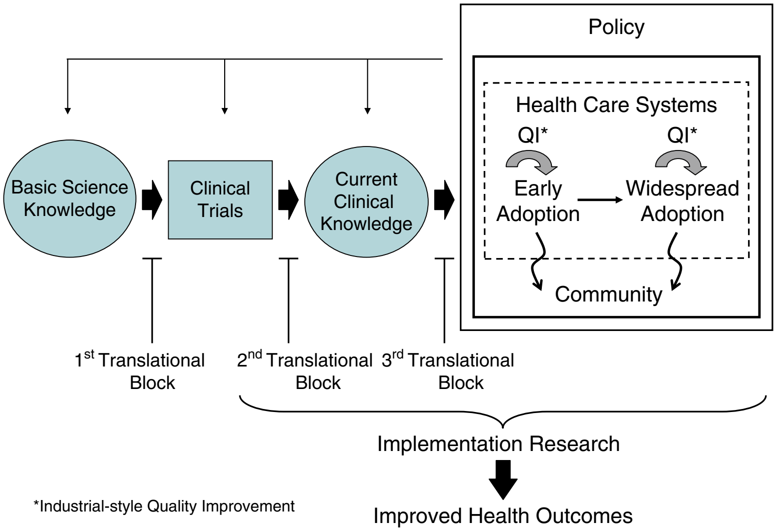


Fig. 13.1 Translational blocks targeted by implementation research

As an underlying source of challenge for the U.S. healthcare system as a whole, health disparities contribute equally as a barrier to overcome in implementation research. Members of minority populations such as African-Americans and Latinos in particular, as well as individuals of low socioeconomic status disproportionately fall victim to markedly dismal health outcomes as compared to their white counterparts [40]. Despite the advancements in health and life expectancy as a country, population specific black-white gaps continue to persist in areas such as access to care, quality of care, chronic disease risk factors, and disease incidence and related mortality [41–43]. These examples of inequity have seeped into multiple tiers of the healthcare system from which implementation research is not exempt.

The rapidly changing U.S. demographics indicate that this very marginalized minority population will soon comprise the majority of the U.S. population. The 2012 U.S. Census reported that 50.4 % of 1-year olds born nationwide were racial/ethnic minorities and that Latinos are the largest and fastest growing ethnic group, currently comprising 16.7 % of the population [42, 44]. At this rate of growth, it is postulated that the U.S. will be a majority minority society by 2050, if not sooner [44].

As such, health disparities are an added dimension to the barriers to overcome in evidence implementation. Implementation science aims to make evidence-based findings work in real world patients. However systems, providers and patients in the real world are much different than what is encountered in the research process. Therefore, studies must be specially tailored to include these specific differences among communities and cultures. Should the health disparities in minority groups continue to persist, they will inevitably impede upon the success of implementation research and impose upon the well-being of the nation.

Finally, to promote the spectrum of research depicted in Fig. 13.1, the 2003 NIH Roadmap acknowledges translational research as an important discipline [45].

In fact, several branches of the NIH now have open funding opportunities for implementation research. The integration of research findings from the molecular to the population level is a priority. The Roadmap seeks to join communities and interdisciplinary academic research centers to translate new discoveries into improved population health [46].

Implementation Research Tools

The tools used to translate clinical evidence into routine practice are varied, and no single tool or combination of tools has proven sufficient or completely effective. Furthermore, it may not be the tool itself but how it is implemented in a system that drives change [47]. In fact, this lack of complete effectiveness spurs implementation research to develop innovative adaptations or combinations of currently available tools [48].

Below, we provide an overview of available tools, which are intended as basic building blocks for future implementation research projects. Although different classification systems have been proposed [49], we arranged these tools by their focus: on the patient, the community, the provider, and the healthcare organization. We acknowledge that this classification is somewhat arbitrary because several implementation tools overlap multiple categories.

Patient-Based Implementation Tools

A growing body of evidence suggests that patients may be successfully 'activated' to improve their own care. For example, a medical assistant may review the medical record with the patient and encourage the patient to ask questions at an upcoming visit with the physician. Patients exposed to such programs had better health outcomes, such as improved glycemic control for those with diabetes [50, 51]. In another study, a health maintenance reminder card presented by the patient to the physician at appointments significantly increased rates of influenza vaccination and cancer screening [52].

Other interventions have taught disease-management and problem solving skills to improve chronic disease outcomes. Teaching patient self-management skills is more effective than passive patient education, and these skills have been shown to improve outcomes and reduce costs for patients with arthritis and asthma [53]. As part of the 'collaborative model,' self-management is encouraged through better interactions between the patient, physician, and health care team. The collaborative model includes: (1) identifying problems from the joint perspective of the patient and clinical care team; (2) targeting problems, setting appropriate goals, and developing action plans together; (3) continuing self-management training and support services for patients; (4) active follow up to reinforce the implementation of the care plan [53].

Community-Based Implementation Tools

The Community Health Advisor (CHA) model has been implemented throughout the world to deliver health messages, promote positive health behavior change, and facilitate access to the health care system [54]. Similarly, the Community Health Worker (CHW) model has been used to engage medically underserved communities on a number of different health issues to help individuals overcome financial, social, political, and cultural barriers to health care [55]. Based on the CHA/CHW models, community members, usually without formal education in the health professions, undergo special training and certification in order to carry out an intervention or research protocol in their local community. CHA/CHW interventions have been used to promote prevention and treatment for a large array of conditions, including cancer, asthma, cardiovascular disease, depression, and diabetes. These programs have also been developed to decrease youth violence and risky sexual behavior, and may be especially relevant for underserved populations and those living in rural areas. Although promising, CHA/CHW interventions often rely on volunteer workers who may be vulnerable to stress and burnout from work overload. Also, intense training and oversight is often required to assure the accuracy of the health messages being transmitted. A review by Swider found limited high-quality evidence that CHA interventions actually improve health outcomes, which is postulated to be a result of the poor quality of the studies. As such, Swider also called for additional rigorous research on the efficacy and underlying mechanisms through which CHA/CHW interventions work [56]. A more recent review commissioned by the Robert Wood Johnson Foundation found that specific CHA interventions may reduce health disparities, particularly for patients with hypertension and diabetes [17].

As a vital community-based implementation tool, it is useful to consider the basic principles of interaction with a community as the foundation for research success. In the city of Lawrence, Massachusetts the Lawrence Research Initiative was created in order to promote community-participatory and community-responsive research [57]. A document to closely guide the research process was created that included:

- The core principles of a partnership approach to research
- Questions for research partnership agreements for researchers and community groups to review
- Steps to building successful research partnerships
- Glossary of research terms in order to develop a common vocabulary that empowers the community to communicate with researchers.

The core principles of partnership as defined in the Lawrence Research Initiative are notions that are applicable to a broad scope of community-based implementation projects [58]. The principles are as follows:

- Research is helpful to community development
- True partnerships between the community and academia make better science

- Researchers and members of the community can and should create good partnerships based on fairness and positive exchanges.
- Partnerships should be based on fair and equitable distribution of resources

For community-based implementation research to be truly successful, it should be rooted in the foundation of good community partnerships and relationships. This is a principle that will hold true across the use of community health workers, community health advisors, and community based participatory research. Equitable partnerships are key.

Provider-Based Implementation Tools

Clinical Guidelines

Clinical guidelines have been defined as 'systematically developed statements to assist practitioners' and patients' decisions about appropriate health care for specific clinical circumstances' [59]. Ideally, guideline development involves a complete review of the relevant literature; however, a Canadian group demonstrated literature reviews were less likely to be completed on more recently developed guidelines [60]. During the last 30 years, guideline dissemination efforts may be suboptimal, leading only to modest improvements in care [60, 61]. However, guideline dissemination alone is not sufficient for implementation of best practices [62].

For many clinical situations encountered today, thousands of evidence-based guidelines and practice recommendations have been published. Such sheer volume often precludes the individual practitioner from implementing all recommendations for every patient. As an example, Boyd et al. noted that if one were to treat a hypothetical 79 year old woman with diabetes, chronic obstructive pulmonary disease (COPD), hypertension, osteoporosis, and osteoarthritis, and follow all recommended guidelines for her multiple co-morbidities, the patient would require 12 medications at a cost of $406 per month [63].

Continuing Medical Education

Continuing medical education (CME), a requirement for ongoing medical licensure, has traditionally relied on text-based, didactic methods to affect clinical knowledge, skills, attitudes, practice patterns, and patient outcomes [64]. However, passive, text-based educational materials and formal CME conferences do not lead to measurable improvements in practice patterns [65, 66]. Rather, CME using interactive techniques which actively engage physicians may have small effects improving practice patterns and patient outcomes [67–70]. Physicians who reflect on their own individual performance may identify areas for improvement and seek CME through multifaceted, self-directed learning opportunities. The use of multiple modalities that promote active learning – such as case-based problem solving – has yielded modest improvements in clinical practice [71]. In general, however, more complex behavioral change likely requires practicing skills similar to traditional quality improvement (QI) [72].

With the advantages of being convenient, flexible, and inexpensive, the Internet has become a useful platform to reach a wider audience for interactive CME, while maintaining an effectiveness comparable to traditional approaches [73]. Fordis et al. conducted a randomized controlled trial comparing live, small-group interactive CME workshops with Internet CME [74]. Both groups focused on cholesterol management. All physicians received didactic instruction, interactive cases with feedback, practice tools and resources, and access to expert advice. Knowledge scores for physicians in the Internet CME group increased more than scores for those in the live CME group. Additionally, the online CME group demonstrated a statistically significant improvement in appropriate drug treatment for high-risk patients. Success of the Internet CME may have been partially driven by the participants' ability to repeatedly return to the website for reinforcement and the ability to structure the learning experience to meet individual needs.

Academic Detailing

Academic detailing relies on site visits to physicians' offices for intense relationship building and one-on-one information delivery. Important components for successful detailing include: (1) assessment of baseline knowledge and motivations for current behavior; (2) articulating clear objectives for education and behavior; (3) gaining credibility with ties to respected organizations through ongoing relationship building; (4) encouraging physicians to actively participate in educational interventions; (5) using graphic representations for educational materials; (6) focusing on a limited number of 'take-home' points; and, (7) supplying positive reinforcement for improved behaviors during follow up [75]. Representatives from pharmaceutical companies have effectively used academic detailing to boost product sales. In a systematic review, academic detailing alone yielded small effects on medication prescribing practices [76].

Opinion Leaders

Several implementation programs have relied on influential colleagues to disseminate evidence-based practices [79]. Opinion leader strategies may include using celebrities, employing people in leadership positions, and asking those doing front-line work to refer 'up the ladder.' In a systematic review, opinion leaders may have a positive effect on evidence-based practice uptake when tested in randomized controlled trials [77].

Physician Audit and Feedback

The utility of audit and feedback hinges on developing credible data-driven summaries of how patient populations are being managed. In theory, such reports may prompt clinicians to reflect on their personal clinical practices and motivate subsequent improvement. Performance feedback may focus on outcomes (such as percentage of patients with diabetes who have achieved glycemic control) or process (such as the percentage of patients with diabetes for whom the physician measured glycemic control). The credibility of performance feedback relies on the ability to capture the many clinical nuances that the physician must consider when delivering care to the individual patient. Because the difficulties in capturing these clinical nuances have not yet been completely surmounted, comparisons of performance to a data-driven, peer-based benchmark may be more appropriate than comparison to an arbitrary standard of perfect performance [80]. A systematic review of randomized trials of audit and feedback studies demonstrated small effects on professional performance. The effect varied by which targeted behavior was chosen. Additionally, the analysis suggests that audit and feedback may be more effective when: the baseline performance is low; the information is provided by a manager or colleague multiple times, communicated in verbal and written formats, and when it is linked to specific goals and an action plan [78].

Organization-Based Implementation Tools

Industrial-Style Quality Improvement

This type of improvement activity originated outside of health care and has acquired such labels as Total Quality Management (TQM) and Continuous Quality Improvement (CQI). These approaches make two fundamental assumptions: (a) that poor outcomes are attributable to system failures, rather than lack of individual effort or individual mistakes, and (b) achieving improvement and excellence, even in the absence of system failures, is possible through iterative cycles of planning, acting, and observing the results. In general, complex systems must have built-in redundancy to function well. If an individual makes a mistake at one point in the system, checks and balances built into other parts of the system may prevent an adverse event. However, as described in the example below, patient safety maybe endangered by simultaneous failure of multiple system components, thus defeating built-in redundancy.

As a simple example, multiple mechanisms should be in place to ensure that incompatible blood products are not given to hospitalized patients. Delivery of the wrong blood type to a patient requires failure at multiple points, including preparation of the blood in the blood bank and administration of the blood by the nurse. Taking such a systems approach stands in stark contrast to blaming individuals, thereby avoiding low morale and reluctance to disclose mistakes.

Improvement activity usually proceeds through a series of ‘plan-do-study-act’ cycles. These cycles emphasize measuring the process of clinical care delivery at the level of the clinical microsystem, which has been previously described. Here, small amounts of data guide the initial improvement process. The process emphasizes small, continuous gains through repeated cycles and does not rely on the statistical significance of the measurements. Although many health care institutions have adopted such methodology based on compelling case studies, additional studies with high-quality experimental methods are still needed [81].

Systems Reengineering

Instead of incremental changes to clinical microsystems, major redesign of the entire system may be undertaken. For example, in the 1990s the Veterans’ Health Administration (VHA) undertook a major reengineering of its health care system, focusing on the improved use of information technology (IT), the measurement and reporting of performance, and the integration of services [82]. By 2000, the VHA had made statistically significant improvements in nine areas, including preventive care, outpatient care (diabetes, hypertension, and depression), and inpatient care (acute myocardial infarction and congestive heart failure). Additionally, the VHA performed better than the fee-for-service Medicare system on 12 of 13 quality measures [82]. Because systems engineering requires changes on such a large scale, little evidence exists about its efficiency and effectiveness in yielding more improvements than smaller changes [3].

With the passage of the Affordable Care Act of 2010, U.S. healthcare systems are making necessary changes to how care is delivered and reimbursed in order to increase the quality of care, reduce costs, and improve patient outcomes. The Center for Medicare & Medicaid Innovation (CMMI) [83] has funded hundreds of demonstration projects to test new models of accountable care organizations, primary care transformation, bundled payments of related services, adoption of best practices, and new care and payment models. The aim is for healthcare systems to implement improvements, and demonstrate positive changes to patient outcomes and cost savings, then CMMI will disseminate the knowledge for quick uptake. Further study of how these changes are implemented as well as program evaluation is needed in the coming years.

Computer-Based Systems

Computer-based systems target links in the process of care delivery that are most prone to human error. Such systems may provide clinical decision support by assisting the clinician with making a diagnosis, choosing among alternative treatments, or deciding upon a particular drug dosage. Other functions may include delivery of clinical reminders and computerized provider order entry (CPOE) [84]. A systematic review documented improvements in time to therapeutic goals, decreases in toxic drug levels and adverse reactions, and shorter hospital stays [85]. However, adverse effects of computer-based systems have also been reported, including increased mortality rates, increased rates of adverse drug reactions, delays in medication administration, increased workload, and new types of errors [86–89]. A systematic review of studies reporting the effect of CPOE on inpatient medication errors also demonstrated mixed results [90], with increases as well as decreases in errors after the introduction of computer-based systems. Therefore, computer-based systems should not be implemented without safeguards to prevent unintended consequences. We need more work to better understand how computer-based systems interact with human users and the complex health care environment and how these interactions affect quality, safety, and outcomes.

Public Report Cards

Public reports on the quality of health care delivered by institutions are proliferating. For example, public reports may focus on risk-adjusted mortality after cardiac surgery or quality at long-term care facilities. In addition, such reports will probably be expanded to include physician groups and individual physicians. Public reports are often promoted under the assumption that the public will use them to choose high-quality providers, thus better enabling a competitive 'medical marketplace.' Although scant evidence links report cards to improved health care [91], report cards may have profound adverse effects: (1) physicians may avoid sicker patients to improve their ratings; (2) physicians may strive to meet the targeted rates for interventions even in situations where intervention is inappropriate; and, (3) physicians may ignore patient preferences and neglect clinical judgment [92]. Even worse, report cards may actually widen gaps in health disparities [92].

The Centers for Medicare & Medicaid Services (CMS) introduced public reporting of quality measures in 2001 [93]. In the last several years, the amount of information accessible to the public has grown tremendously. On the CMS website, the public can see data from hospitals, inpatient rehabilitation facilities, long term care hospitals, nursing homes, and home health agencies.

Pay-for-Performance (P4P)

Currently, there is mounting pressure to tie reimbursement for health care services to quality measurement. Although allowing market forces to freely operate through P4P reimbursement may seem logical, systematic reviews have not yielded conclusive results. Because not everything that is important is currently measured, linking reimbursement to measured quality may divert attention from important, but unmeasured aspects of care (i.e., 'spotlight' effect). As with public reporting, P4P may actually widen health disparities, although empiric data are lacking.

To date, evidence for the effectiveness of P4P in improving the delivery of health care is uncertain [94]. One study found that when implemented in physician practice groups, P4P produced improvements for those with higher baseline performance but had minimal effect on the lowest performers [95]. Glickman et al. found hospitals voluntarily participating in the P4P initiative for myocardial infarction did not show appreciable improvement [96]. A recent study found that hospitals participating in P4P and public reporting programs sponsored by CMS had slightly greater improvements in quality than those only participating in the public reporting program [97]. Several ongoing studies may soon deliver new insights about P4P.

Designing Implementation Research Studies

Because the implementation science base is still emerging, researchers have at their disposal an array of tools that are variously effective, depending upon the patient population and delivery setting. Moving beyond the tools described above, is the need to develop innovative adaptations and approaches to bridge the gap between clinical knowledge and health care practice. It is necessary to test the effectiveness of these new approaches with rigorous scientific methods to avoid adverse consequences from the wide-spread dissemination and adoption of unproven interventions [30]. Therefore, in the remainder of this chapter, we discuss the critical design elements for implementation randomized controlled trials, followed by an example of an implementation research study.

Overview of Implementation Research Study Design

Randomized designs for implementation research, somewhat analogous to the traditional clinical trial, allow causal inference and offer protection from measured and unmeasured confounding (See Chap. 3) [48]. As described below in more detail, such designs include an active intervention, random allocation to a comparison or intervention group, and blinded assessment of objective endpoints. Although many of the same principals are involved in clinical RCTs and implementation RCTs, these will be reviewed with emphasis on some of the differences between the two.

Falling lower in the hierarchy of evidence, implementation studies may use other non-randomized or controlled designs. For example, a research team may observe a single group for changes in health care delivery or patient outcomes before and after intervention implementation. In this case, the observed changes may result from multiple factors not associated with the intervention. Secular trends may produce broad, population-based changes, independent of the intervention under study. Without a comparison group, secular trends may be confused with intervention effects [48]. Interrupted time-series designs, with data collected from multiple points in time before and after the intervention, can better account for secular trends.

Table 13.2 Key components for implementation research proposals

1. Gap in care or quality of care identified
2. Evidence-based of intervention demonstrated
3. Conceptual model and theory justified
4. Stakeholders' priorities recognized and engagement proven
5. Setting's readiness to adopt intervention articulated
6. Implementation strategy and process defined and justified
7. Team experience with the setting, treatment, and implementation process demonstrated
8. Research design, methods, and contingency plans are feasible
9. Measurement and analysis detailed and scientifically sound
10. Policy and funding environment aligned

Adapted from Proctor et al. [33]

In addition to confounding from secular trends, uncontrolled study designs are susceptible to other 'non-interventional' aspects of the intervention. For example, an intervention may bestow more attention on patients or clinicians through data collection, leading to self-reported improvement through placebo-like effects. Comparison groups, even without randomization, offer important protection against secular trends and placebo-like effects. Non-randomized allocation to intervention and comparison groups does not assure that both groups are similar in all important characteristics. Matched study designs may balance study groups for a limited number of measured characteristics. In contrast, successfully implemented randomization equalizes recognized and unrecognized confounders across study groups and is, therefore, essential for cause-and-effect inference.

In summary, limitations of study designs without randomization or a comparison group include difficulty establishing causality, confounding, bias, and spurious associations from multiple comparisons [29]. Although such studies are generally considered to be lower within the evidence hierarchy, they may provide useful information when randomized controlled trials (RCTs) are not feasible or generate important hypotheses for subsequent testing with more rigorous study designs. We focus the remainder of this chapter on key RCTs for implementation research, in particular the cluster RCT – where clusters of individuals (groups) are randomized [98] rather than individuals. Due to the complex design, we strongly recommend that investigators obtain expert consultation with methodologists and statisticians early during the planning stages.

Other study designs applicable for implementation research and quality improvement projects are reviewed elsewhere [99, 100]. Proctor et al. reviews funding mechanisms and provides key ingredients for implementation research proposals (Table 13.2) [33]. Competencies for trainees of implementation and dissemination research are also available elsewhere [101].

Implementation Randomized Controlled Trials

Many principles for the design of high-quality, traditional RCTs discussed elsewhere in this book also apply to implementation research. As a discussion guide, our approach parallels the Consolidated Standards of Reporting Trials (CONSORT), which were designed to encourage high-quality clinical randomized trials and promote a uniform reporting style. The CONSORT criteria emphasize the ability to understand the flow of all actual and potential research participants through the experimental design. Although originally designed for the traditional or 'parallel' clinical trial [102, 103], the CONSORT criteria were subsequently modified for the cluster RCT [104, 105].

We refer the reader to specific example of an implementation randomized trial illustrating the formative development of one of the outcomes [106], challenges and barriers with recruitment [107], main outcome, [108] and secondary outcomes [109] (ClinicalTrials.gov Identifier: NCT00403091; Available at: http://clinicaltrials.gov).

Participants and Recruitment

In contrast to the randomized clinical trial where patients are the unit of intervention and analysis, implementation randomized trials and interventions have a broader reach. For example, key participants in implementation RCTs may be doctors, patients, clinics, or hospitals, or hospital wards. Because implementation research is conducted in the 'real world' and often seeks to engage busy clinicians, systems, and patients who are otherwise overwhelmed with their usual activities, recruitment may be particularly difficult. Therefore, recruitment protocols for implementation research demand careful consideration and may require a dedicated recruitment and retention team that is specific to the target population. Often multiple approaches (e.g., word of mouth, e-mail, phone, fax, personal contacts, or lists from professional organizations) must be pursued, and still the desired number of participants may not be reached. This is a particular challenge as it pertains to recruiting individuals and engaging systems that cater to marginalized minority populations.

African Americans and Latinos continue to bear an unequal burden of disease. Individuals from these populations are underrepresented in implementation research. To reach wide applicability, a diverse pool of participants in research studies is necessary. However, racial and ethnic minorities remain underrepresented in research participation. For example, less than one-third of those enrolled in research studies sponsored by the National Institutes of Health (NIH) are minorities [110, 111],– African Americans comprising 12.6 % and Latinos 7.5 % [111].

Minorities have often been underrepresented in traditional clinical research studies for several reasons. Researchers and participants often do not share common cultural perspectives, which may lead to lack of trust. Moreover, limited resources, such as low levels of income, education, health insurance, social integration, and health literacy, may also preclude participation in research studies studies [112]. In

Table 13.3 Solutions to commonly faced barriers to minority recruitment

Barrier encountered	Offered solution to overcome barrier
Lack of public awareness/community participation	Creation of culturally sensitive, targeted marketing for recruitment
Underrepresentation of minorities in a population sample	Oversampling of targeted minority population
Limited research literacy of target population	Creation of culturally and linguistically competent study materials. This may include language translation, or use of vernacular terms specific to a particular community
Unfamiliarity with community/where to find target population	Dispersal of recruitment materials to areas of broad attendance such as: mass transportation, radio stations, grocery stores
Researchers may neglect to offer research studies to individuals from underrepresented groups	Researchers should offer research study participation to all, negating preconceived notions about who may or may not have an interest in participating

addition, the history of racism in the U.S. and particularly in medical research and clinical care, has contributed to deep suspicion among minority communities about the motives of the medical system [113–115].

Low research participation from communities of color stems directly from these historical inequities and power imbalances that have created a lack of trust between community and academic medical institutions.

Within the past two decades, a series of nationwide mandates for federally funded research have been created in order to directly address the concerns of distrust in these populations including: the NIH Revitalization Act created in 1993 and updated in 2001 mandating the inclusion of women and minorities in clinical trials [41, 116], the 1997 Federal and Drug Administration (FDA) Modernization Act providing strict requirements on the standardization of data collection on racial/ethnic minority groups in clinical trials [116], and the Centers for Medicare and Medicaid Services (CMS) authorization of routine care costs for Medicare beneficiaries who are participants in clinical trials in 2000 [116].

Despite these mandates, challenges in the recruitment of minorities still exist. Chapter 8 of this textbook offers additional insight on broad recruitment strategies for implementation research. Table 13.3 offers some solutions to these commonly faced barriers [117].

Human Subjects

Review and approval of implementation studies by an institutional review board (IRB) is necessary. Often, the research protocol may pose minimal danger to participants and the review may be conducted under an expedited protocol. We refer the reader to more detailed reviews on this topic [118–120]. Randomization, intent to publish study findings, or present at scientific conferences places the work in the research domain. Although usual local quality improvement activities, which are important in health care, do not require IRB approval, the addition of a rigorous design for implementation research does require review.

Investigators designing cluster RCTs must carefully consider the ethical issues that arise when consent occurs at the cluster level with subsequent enrollment of participants within the cluster. If the target of the research is clearly the clinician, informed consent may often be waived for the patient. For studies that focus on the clinician but collect outcomes from medical record review or administrative patient records, the researchers may consider applying for a waiver of informed patient consent. Such waivers are especially reasonable when a large volume of patient records would make patient informed consent impractical. Implementation research usually generates personally identifiable health information, which may be subject to the Health Insurance Portability and Accountability Act (HIPAA). Waiver of HIPAA consent by the patient may often be obtained based on requirements similar to waiver of informed patient consent. Finally, it may be necessary to obtain consent from both patients and providers if the intervention targets both populations.

Investigators should develop detailed plans to protect the security and confidentiality of study data. Data should be housed in physically secured locations with strong logical protection, such as password protection and encrypted files. Access to study data should be only on a 'need-to-know' basis. Participant identifiers should be maintained only as necessary for data quality control and linkage. Patients and clinicians should be assured that personal information will not be revealed in publications or presentations. Data integrity should also be protected with detailed protocols for verification and cleaning, which are beyond the scope of this chapter [121].

We agree with the International Committee of Medical Journal Editors (ICMJE) that descriptions of all randomized clinical trials should be deposited in publically available registries before recruitment begins [122]. The ICJME includes interventions focusing on process-of-care within the rubric of clinical trials. Trial registries guard against the well-recognized bias that negative studies are less likely to be published than positive studies. Negative publication bias may significantly limit meta-analytic studies, leading to the false conclusion that ineffective interventions are actually effective. Registries also increase the likelihood that participation in clinical trials will promote the public good, even if the study is negative. Although the template is not customized for implementation research, one such registry may be found at http://clinicaltrials.gov.

Intervention Design

Based on the concepts described earlier in this chapter, the design of the intervention is often guided using a formative-evaluation process [123, 124]. Formative evaluation incorporates input from end users and stakeholders to refine an intervention during the early stages of development. It is critical that investigators carefully explore and understand the need of those who will be affected by the intervention. For example, the design can be guided by focus groups or nominal group technique [125–128]. Glasgow et al. [37] recommend key features to include in the content design:

- barrier analysis
- integration of multiple types of evidence
- adoption of practical trials that address clinician concerns
- investigation of multiple outcomes, generalizability, and contextual factors
- design of multilevel programs using systems and social networking models mindful of the integration of the study's components and levels, and
- adaptation of program to local needs and ongoing issues.

For example, for an internet-delivered intervention for physicians important features to consider include [129]:

- needs assessment from office practice data
- multimodal strategies
- modular design with multiple parts
- clinical cases for contextual learning
- tailoring intervention based on individual responses
- interactivity with the learner
- audit and feedback
- evidence-based content
- established credibility of organization providing website and funding entity
- patient education resources
- high level of usability, and
- accessibility to the Internet site despite limited bandwidth.

Comparison Group

In behavioral research, it is often appropriate to randomize participants to either an active intervention versus an attention control. The attention control – in contrast to 'placebo' or no intervention- accounts for changes in behavior attributable to social exposure when participants receive services and attention from study personnel [130]. Positive social interactions may create expectations for positive outcomes, potentially confounding intervention effects collected through such methods as self-report. However, the precise implementation of attention controls may be difficult [131].

In our experience, clinicians and communities may be reluctant to enter a study with the possibility of being randomized to a group with no apparent benefit. This problem may be compounded by intensive procedures needed for data collection, regardless of the study group. To overcome such barriers, investigators may offer to open the intervention to the comparison group at the close of the study. Alternatively, study design might more formally incorporate a delayed intervention or test two variations of an active intervention.

Blinding

'Blinding' is important to decrease bias in outcome ascertainment (similar to randomized clinical trials). Study personnel who perform outcome assessment should be unaware of whether an individual participant has been assigned to the intervention or comparison group. For example, it may be necessary to blind those doing patient examinations, those performing medical record abstraction, or those administering patient, physician, or organizational surveys. When participants are blinded to the allocation arm, the study is single-blinded. If those delivering the intervention and collecting the outcomes are blinded as well, then the study is double-blinded. If the analysts are unaware of the assignments, then the study is triple-blinded. For implementation research, it is often not feasible to conceal study allocation from the research team.

Units of Intervention, Randomization, and Analysis

Investigators planning an implementation randomized trial must carefully consider the units of study assignment for intervention, randomization, and analysis. Examples of units of intervention are patients, physicians, nurses, clinics, hospitals, hospital wards, among others. Within any given study, the unit level may vary across components, meaning that the analysis plan must account for the clustered nature of the outcome data.

For example, consider a study of a patient-based intervention that will be implemented through a group of affiliated multi-physician clinics. 'Contamination' could arise from physicians learning about the intervention and then exposing comparison patients to part of the intervention. Therefore, for this particular study, the investigators may choose to randomize at the physician level to avoid contamination. Thus, all patients assigned to a given physician will be allocated to the same condition: intervention or comparison.

In practice, the threat of contamination may be more perceived than real, depending upon the exact nature of the intervention and study setting. When present, contamination decreases the precision with which the intervention effect will be measured and increases the risk of a Type II error. As an alternative to cluster-based randomization to overcome contamination, the sample size could be increased [122].

Measurement and Outcomes

In implementation research, the science of determining an approach to define the measures to obtain and the specific outcomes is rapidly evolving [132]. Concepts of treatment integrity utilized in the traditional randomized controlled trial also apply to implementation research, also known as treatment fidelity. In addition, concepts are also applicable for the assessment of *external validity* – the applicability of the findings in other settings.

The use of a systematic strategy allows the implementation researcher to plan ahead and define the measures and relevant outcomes. The ultimate goal is to have a strong foundation for formative and summative evaluations utilizing quantitative and qualitative methods. Similar in importance as knowing whether an intervention worked (or did not) is to understanding '*how*' the intervention worked (or did not).

Research that uses a mixed methods (or multimethods) approach is suitable to understand problems from multiple perspectives and contextualize information [133]. Mixed methods research is defined as "the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration" [134].

Strategies are available to design, evaluate, and report implementation research studies. A systematic review and a book are available elsewhere; [135, 136] in this chapter, we briefly review the following:

- Reach, Effectiveness, Adoption, Implementation, and Maintenance (***RE-AIM***)
- Pragmatic-Explanatory Continuum Indicator Summary (**PRECIS**)
- Predisposing, Reinforcing, and Enabling Constructs in Ecosystem Diagnosis and Evaluation (***PRECEDE***)-Policy, Resourcing, and Organization for Educational and Environmental Development (***PROCEDE***).
- Realist Evaluation

The '***RE-AIM***' approach is particularly helpful to evaluate the public health impact of interventions [37, 137–142]:

- ***R***each – the intended target population, study's reach and representativeness, participants and setting – 'How many participate?'
- ***E***ffectiveness – the magnitude of intervention effect, adverse outcomes, and costs – 'Does it work in usual settings?'
- ***A***doption – use by the target audience – 'How many use it?'
- ***I***mplementation – the consistency of use, costs, and adaptations made during delivery – 'Is it used as intended?'
- ***M***aintenance – the intervention's long-term effects, sustainability, and attrition rates - 'Is it sustained over time?'

The Pragmatic-Explanatory Continuum Indicator Summary (**PRECIS**) originated from a Canadian and European initiative to promote trials in developing and middle-income countries [143]. Within this context, a *pragmatic trial* seeks to answer the question, "Does an intervention work in usual settings under usual conditions?;" a pragmatic trial tests the effectiveness of the intervention and informs decision makers [144]. An *explanatory trial* seeks to answer the question, "Does an intervention work in research settings?;" an explanatory trial tests the efficacy of an intervention [145, 146].

The visual representation of measures among the ten domains described in PRECIS allows scientists and community in general understand the applicability of the interventions. The ten domains include: (1) participant eligibility criteria, (2) experimental intervention flexibility, (3) practitioner expertise (experimental), (4) comparison intervention, (5) practitioner expertise (comparison), (6) follow-up intensity, (7) primary trial outcome, (8) participant compliance, (9) practitioner adherence, and (10) analysis of primary outcome. For a more detailed discussion, the reader is referred elsewhere [143, 147].

First proposed in 1974, the ***PRECEDE-PROCEED*** is an approach to assess the effects of health programs and health education applicable for implementation research [148–150]. The realist approach offers a quantitative and qualitative model for synthesis of the effects of complex programs that are intimately related to the contextual factors where the program is developed and evaluated [151–153]. A realist approach addresses some of the short-comings of purely quantitative methods for program evaluation [152].

Reviews are available to guide the design, measurement, and reporting of implementation research studies [132, 142, 144, 154]. The addition of information about the context, protocol implementation, and generalizability – among other characteristics – are enhancements to the CONSORT reporting guidelines for the traditional efficacy study [142, 154].

Approaches to Randomization

Randomization, also described elsewhere in this book, is a procedure to assure that study units are allocated to the study conditions according to chance alone. The specific approach to randomization is described as 'sequence generation' and may include matching or stratification [104]. *Allocation concealment* is a 'technique used to prevent selection bias by concealing the allocation sequence from those assigning participants to intervention groups, until the moment of assignment, the purpose of which is to prevent researchers from influencing which participants are assigned to a given group [102, 103]. The concealment may be simply based on a coded list of randomly ordered study groups created by a statistician who is not a member of the intervention team. After enrollment, each participant is assigned to a study group based on the sequence in the list.

For cluster-randomized trials, the assignment of individuals to a study group is determined at the level of the cluster, which increases the opportunity for selection bias from failed concealment. For example, consider a cluster RCT where randomization occurs at the physician level with subsequent enrollment of patients with diabetes from the physicians' practice. Depending upon the nature of the intervention, physicians may be able to determine their randomization group. If the randomized physician also recruits patients for the study, this knowledge of the randomization group may lead to biased patient selection. An 'attention control' comparison group described above would also decrease the chances of the physician to discover the assignment.

Successful randomization ensures balanced characteristics at the unit of randomization, and larger numbers of randomized units increase the chance of successful randomization. Investigators should be aware that for cluster RCTs, successful randomization does not ensure balanced characteristics at units below the level of randomization [155]. Again, consider the illustration above where randomization occurs at the physician level. Although this design may produce intervention and comparison groups that are balanced based on physician characteristics, there may be important imbalances in patient characteristics, decreasing the power of randomization. To guard against imbalances of lower-level units in cluster randomized trials, investigators might consider stratifying or matching on a limited number of critical characteristics [156]. Alternatively, imbalances may require statistical adjustment at the point of analysis after the study has been completed. Decisions about matched study designs for cluster randomized trials are complex and beyond the scope of this chapter.

Intent-to-Treat

As with the traditional clinical randomized trial, the primary analysis for an implementation randomized trial should test hypotheses specified *a priori* and should follow intent-to-treat principles [157]. With the intent-to-treat approach, all units are analyzed with the group to which they were originally randomized, regardless of whether the units are subsequently exposed to the intervention (i.e., cross over). For example, in a randomized trial of an Internet-based continuing medical education (CME) intervention for physicians, outcomes for all physicians randomized to the intervention group must be analyzed as part of the intervention group, regardless of whether the physician visited the Internet site. Intent-to-treat protocols preserve the power of randomization by protecting against bias resulting from differential participation or cross-over among intervention units with a greater or lesser propensity for success.

Unfortunately, participants lost to follow up may generate no data for analysis. As with violation of the intent-to-treat principle, loss to follow up may reduce the power of randomization. Although complete follow up is desirable, it is usually not obtainable. Many scientists hold that for clinical trials, loss to follow up of greater than 20 % introduces severe potential for bias [158]. Therefore, many study designs include run-in phases before randomization. From the perspective of internal validity, it is better to exclude participants before randomization than have participants lost to follow up, cross between study groups, or become non-adherent to intervention protocols after randomization. For example, in the study of Internet-based CME described above, physicians might be required to demonstrate a willingness to engage in Internet learning and submit data for study evaluation before randomization. According to the CONSORT criteria for group randomized trials, investigators must carefully account for all individuals and clusters that were screened or randomized [104].

Retention, Special Populations

Retention of research participants is a challenging issue in research, and can be of particular concern when working with vulnerable and underserved populations. The broad goal of implementation science is to translate evidence-based practice into real world application. Specific to this goal, there is an overarching need to target and tailor implementation to the specific system, providers, and patients that exist in a given community. Methods that to date have been typically implemented at the system and provider level within academic health settings, and included predominately homogenous white patient participation, will not translate well in more diverse, community driven settings.

Lack of retention and loss to follow up can be a barrier in implementation research, especially for projects concerning health disparities in minorities. In the recently conducted Healthy Aging in Neighborhoods of Diversity Across the Life Span (HANDLS) study, Ejogu et al. present a multifaceted approach to recruitment and specifically retention strategies for minority and low socioeconomic status (SES) participants [159]. In this 20-year longitudinal examination of how race and SES influence the development of age-related health disparities, the investigators created a multifactorial recruitment and retention strategy that targeted known barriers and identified those unique to the study's urban environment [159]. Through this approach, they were able to recruit over 3,700 participants, of whom 59 % were African American with a 75 % baseline completion [159]. The success of the HANDLS investigation relied primarily on the emphasis of the community-based platform to alleviate many of the barriers that might exclude this key population from participation and retention.

Underrepresented minorities are a special population whose participation in implementation research holds promise in revealing methods to reduce health disparities [160–162]. It may be difficult to ascertain the true population benefit or effectiveness of an intervention if a significant proportion of its participants are lost to follow-up. While Chap. 8 of this textbook is solely dedicated to addressing broad strategies for recruitment and retention, this section offers specific insight into retention issues pertaining to the participation of minorities. As an overarching challenge between the health care system and minority communities, the establishment of trust continuously strikes a chord as a key necessity in retaining the attention and participation of this population in implementation research [160, 163]. Yancey et al. suggest some targeted approaches to decreasing participant loss specifically in underrepresented and minority groups [160–162, 164–166]:

- Intensive follow-up and contact with subjects
- Retain interviewers, field staff, and study staff over time
- Involve staff from the targeted community
- Provide social support and offering accessible locations for study visits and/or data collection
- Ensure timely incentive payments and accessibility of project staff
- Encourage study staff's knowledge of community dynamics and project leadership/staff visibility and involvement in the community.

Statistical Analysis

Statistical analysis for cluster RCTs is a vast, technical topic that falls largely beyond the domain of the basic introduction provided in this book. However, an example will illustrate some important principles. More specifically, consider the previous illustration in which physicians are randomized to an intervention or comparison group, with patients being subsequently enrolled and assigned to the same study condition as their physician. To conduct the analysis at the physician level, the investigators might simply compare the mean post-intervention outcomes for the two study groups. However, this approach leads to loss of statistical power, because the number of physicians randomized will be less than the number of patients included in the study. Alternatively, the investigators could plan a patient-level analysis that appropriately considers the clustering of patients within physicians. The investigators could also collect outcomes for intervention and comparison patients before and after intervention implementation. Generalized estimation equations could then be used to compare the change in study endpoints over time for the intervention versus comparison group. Here, the main study effect will be reflected by a group-time interaction variable included in the multivariable model. This approach uses a marginal, population-averaged model to account for clustered observations and potentially adjust for observed imbalances in the study groups. Alternatively, the analyst may use a cluster-specific (or conditional) approach that directly incorporates random effects. Murray reviewed the evolving science and controversies surrounding the analysis of group-randomized trials [156].

Although the main analysis should follow intent-to-treat principles as described above, most implementation randomized trials include a range of secondary analyses. Such secondary analyses may yield important findings, but they do not carry the power of cause-and-effect inference. ‘Per-protocol’ or ‘compliers only’ analyses may address the impact of the intervention among those who are sufficiently exposed or may examine dose-response relationships between intervention exposure and outcomes. Mediation analysis using a series of staged regression models may investigate mechanisms through which an intervention leads to a positive study effect [167, 168].

Sample Size Calculations

The investigator must determine the number of participants necessary to detect a meaningful difference in study endpoints between the intervention and comparison groups, i.e., the power of the study. Typically, a power of 80 % is considered adequate to decrease the likelihood of a false negative result. If an intervention is sustained over an extended period of time, the investigators may wish to test specifically for effect decay, perhaps with a time-trend analysis. Such a hypothesis of no difference demands a special approach to power calculation. Sample size calculations for traditional randomized trials are discussed elsewhere in this book (see Chap. 15).

The analysis for an implementation randomized trial may be at a lower level than the unit of randomization. Under these circumstances, the power calculations must account for the clustering of participants within upper-level units, such as the clustering of patients within physicians from the example above. Failure to account for the hierarchical data structure may inflate the observed statistical significance and increase the likelihood of a false positive finding [169].

Several approaches to accounting for the clustering of, say, patients within physicians from the above example, rely on the intra-class correlation coefficient (ICC). The ICC is the ratio of the between-cluster variance to the total sample variance (between clusters + within cluster). In this example, the ICC would be a measure of how 'alike' patient outcomes were within the physician clusters. If the ICC is 1, the outcomes for all patients clustered within a given physician are identical. If the ICC is 0, clustering within physicians is not related to patient outcomes [170]. In other words, with an ICC of 1, adding additional patients provides no additional information. Therefore, as the ICC increases, one must increase the sample size to retain the same power. For $0 < \text{ICC} < 1$, increasing the number of patients will increase study power less than increasing the number of physicians. Typical values for ICCs range from 0.01 to 0.50 [171].

Although the topic of power calculations for group randomized trials is vast and largely beyond the scope of this book, Donner provides a straight-forward framework for simple situations [169]. Taking this approach, the analyst first calculates an unadjusted sample size (N_{un}) using approaches identical to those described elsewhere in this book for the traditional randomized clinical trial. Next, the analyst calculates a sample inflation factor (IF) that is used to derive a cluster-adjusted sample size (N_{adj}). Then:

$$\text{IF} = \left[1 + (m - 1)\rho\right] \text{ and}$$

$$N_{adj} = (N_{un}) * \text{IF},$$

where m is the number of study units per cluster, and ρ is the ICC.

Situational Analysis and External Validity

Because implementation randomized trials occur in a 'real-word' setting, we place special emphasis on understanding and reporting of context. In contrast to the traditional randomized clinical trial, the study setting for the implementation trial is an integral part of the study design. To address the importance of context in implementation research, Davidoff and Batalden promote the concept of situational analysis for quality improvement studies [81]. We believe that many of these principles are relevant to the implementation randomized trial. For example, published reports for implementation research should include specific details about the clinic setting, patient population, prior experience with system change, and how the context contributed to understanding the problem for which the study was designed. In addition, specialized approaches to economic evaluation provide additional important context for interpreting the results from implementation trials [172].

Because implementation research often focuses on dissemination to large populations, external validity, or generalizability, acquires special importance. One must consider how study findings are applicable to other patients, doctors, clinics, or geographic locations.

Summary

Implementation research bridges the gap between scientific knowledge and its application to daily practice with the overall purpose of improving the health of individuals and populations. To advance the science of implementation research, the Institute of Medicine published findings from the Forum on the Science of Health Care Quality Improvement and Implementation in 2007 [173] and the Veterans' Health Administration sponsored a state-of-the-art (SOTA) conference in 2004 [3]. Together, these documents summarized current knowledge, identified barriers to implementation research, and defined strategies to overcome these barriers. Given the well-documented quality and safety problems of our health care system despite the vast resources invested in the biomedical sciences, we need to promote interest in implementation research, an emerging scientific discipline focused on improving health care for all, regardless of geography, socioeconomic status, race, or ethnicity.

Acknowledgement We thank Winter Williams, Kierstin Leslie, and Natalie Wilson for critically reviewing a prior version of this chapter.

Resources

Selected Journals That Publish Implementation Research

- *Annals of Internal Medicine*
- *BMJ Quality and Safety in Health Care*
- *Implementation Science*
- *JAMA*
- *Journal of General Internal Medicine*
- *Journal of Hospital Medicine*
- *Medical Care*
- *Pediatrics*
- *The Joint Commission Journal on Quality and Patient Safety*

Selected Checklists and Reporting Guidelines

- Standards for Quality Improvement Reporting Excellence (SQUIRE)
  - SQUIRE are guidelines for publishing quality improvement interventions.
  - The guidelines provide specific details to be addressed in each section of manuscripts that report quality improvement interventions.
  - http://squire-statement.org/
  - Davidoff F, Batalden P. Toward stronger evidence on quality improvement. Draft publication guidelines: the beginning of a consensus project. Qual Saf Health Care 2005;14:319–25.
- Enhancing the Quality and Transparency of Health Research (EQUATOR)
  - EQUATOR is an international initiative that seeks to improve the quality of scientific reporting.
  - This initiative includes statements about reporting for a range of experimental and observational study types, including randomized trials, group randomized trials, behavioral trials, and quality interventions. It also provides education and training on the use of reporting guidelines.
  - http://www.equator-network.org
- Consolidated Standards of Reporting Trials (CONSORT)
  - This initiative focuses on design and reporting standards for randomized controlled trials (RCTs) in health care.
  - Although originally designed for the traditional 'parallel' randomized clinical trial, the CONSORT criteria have been extended to include cluster RCTs and behavioral RCTs.
  - http://www.consort-statement.org/
- Workgroup for Intervention Development and Evaluation Research (WIDER)
  - This checklist is useful in reporting the quality of behavioral change intervention studies.
  - http://interventiondesign.co.uk/wp-content/uploads/2009/02/wider-recommendations.pdf

Selected Resources for Intervention Design

- Agency for Healthcare Research and Quality (AHRQ) website for clinicians and providers
  - The Effective Health Care (EHC) Program invites clinicians to join networks that promote patient-centered outcomes research. http://www.ahrq.gov/professionals/clinicians-providers/

  - Closing the Quality Gap: Revisiting the State of the Science Quality Improvement Interventions to Address Health Disparities: http://www.ahrq.gov/legacy/clinic/tp/gapdisptp.htm
  - Morbidity & Mortality Reviews on the Web. Education site with cases, commentaries, and reviews. http://webmm.ahrq.gov.

- Patient-Centered Outcomes Research Institute: http://www.pcori.org/
- Patient Reported Outcomes Measurement Information System (PROMIS): a National Institutes of Health (NIH) funded system of patient–reported assessment tools to health status. http://www.nihpromis.org/default
- The National Guideline Clearinghouse is a database of evidence-based practice guidelines available to the public. http://www.guideline.gov
- Veterans' Administration Quality Enhancement Research Initiative (QUERI) Implementation Guides
  - The QUERI Implementation Guide is a three-part series focusing on practical issues for designing and conducing implementation research.
  - The guide includes material on conceptual models, diagnosing performance gaps, developing interventions, evaluating implementation research, lessons learned from prior QUERI projects, tools and toolkits, as well as many resources.
  - http://www.queri.research.va.gov/implementation/
- Finding Answers
  - This program is sponsored by the Robert Wood Johnson Foundation to develop interventions for eliminating racial/ethnic disparities in health care.
  - The Finding Answers Intervention Research (FAIR) database includes 388 summaries of journal articles from 11 systematic reviews of interventions to decrease racial/ethnic disparities for many commonly encountered diseases, such as diabetes and hypertension. Interventions based on cultural leverage and performance-based reimbursement are also included.
  - http://www.solvingdisparities.org/
- National Center for Cultural Competence
  - This center is sponsored by Georgetown University and offers several implementation tools, manuscripts, and policy statements for organizations, clinicians, and consumers.
  - The Internet site has a section describing 'promising practices' which may be particularly useful in designing new interventions.
  - http://nccc.georgetown.edu/
- Clinical Microsystems
  - The Dartmouth Institute for Health Policy and Clinical Practice maintains this website that offers tools for improving clinical microsystems.
  - Most tools are generally available to the public at no cost.

  - The Clinical Microsystems Action Guide (under the materials, workbooks tab) may be particularly useful for designing new interventions.
  - http://www.clinicalmicrosystem.org/

- Institute for Healthcare Improvement
  - This not-for-profit organization maintains an Internet site that contains several tools for improving the quality, safety, and efficiency of health care. Many tools are publically available at no cost.
  - White papers describing the 'Breakthrough Series' may be particularly useful for those developing new interventions.
  - http://www.ihi.org

Selected Resources for Implementing and Disseminating Quality Improvement

- Splaine, M. E., Dolansky, M. A., Patrician, P. A., Estrada, C. A. Editors. Oakbrook: Joint Commission Resources. Practice-based Learning and Improvement: A Clinical Improvement Action Guide, 3rd Edition. 2012.
  - Authors explain proven methods for integrating the core competency of practice-based learning and improvement (PBLI) into daily clinical work. Practical tools are described for health professionals working on quality improvement.
- Ogrinc GS, Headrick LA, Moore SM, Barton AJ, Dolansky MA, Madigosky WS. Oakbrook: Joint Commission Resources. Fundamentals of Health Care Improvement: A Guide to Improving Your Patients' Care. 2nd Edition. 2012.
  - The book provides a single source for nursing students, medical students, and resident physicians to learn and practice the basics of QI.
- Brownson RC, Colditz GA, Proctor EK. Dissemination and Implementation Research in Health: Translating Science to Practice. Oxford Scholarship. 2012.
  - The authors provide a comprehensive roadmap for implementation research.

Selected Training Programs

- Veterans Affairs Quality Scholar Fellowship Program
  - A two-year inter-professional education program that offer scholars opportunities to become leaders by applying knowledge and methods of health care improvement to the care of veteran, innovate and continually improve health

care, teach health professionals about health care improvement, perform research and develop new knowledge for the ongoing improvement of the quality and value of health care services.
– http://www.vaqs.org

- Quality and Safety Education for Nurses (QSEN) Institute
  – The Institute offers comprehensive, competency based resources to empower nurses with knowledge, skills, and attitudes to improve quality and safety across the healthcare system.
  – http://qsen.org/
- Training in Dissemination and Implementation Research in Health (TIDIRH).
  – This training is sponsored by the NIH's Office of Behavioral and Social Sciences Research and the U.S. Department of Veterans Affairs.
  – See Meissner et al. Implement Science 2013 Jan 24;8:12.
  – http://conferences.thehillgroup.com/OBSSRinstitutes/TIDIRH2013/
- VA Enhancing Implementation Science in VA Cyber Seminar
  – http://www.queri.research.va.gov/meetings/eis

References

1. Berwick DM. Disseminating innovations in health care. JAMA. 2003;289:1969–75.
2. Lenfant C. Shattuck lecture – clinical research to clinical practice – lost in translation? N Engl J Med. 2003;349:868–74.
3. Kiefe CI, Sales A. A state-of-the-art conference on implementing evidence in health care. Reasons and recommendations. J Gen Intern Med. 2006;21 Suppl 2:S67–70.
4. Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. BMJ. 2012;344:d7292. doi:10.1136/bmj.d7292.
5. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. N Engl J Med. 2003;348:2635–45.
6. National Institute of Health (NIH). (2013, January 9). Dissemination and implementation research in Health (R01). Available at: http://grants.nih.gov/grants/guide/pa-files/PAR-13-055.html. Accessed 9 May 2013.
7. Kiefe CI, Safford M, Allison JJ. Forces influencing the care of complex patients: a framework. In: Academy health annual meeting 2007. Orlando; 2007.
8. Rubenstein LV, Pugh J. Strategies for promoting organizational and practice change by advancing implementation research. J Gen Intern Med. 2006;21 Suppl 2:S58–64.
9. Canadian Institutes of Health Research. More about Knowledge Translation at CIHR. Canadian Institutes of Health Research; 2013. Available at: http://www.cihr-irsc.gc.ca/e/39033.html. Accessed 9 May 2013.
10. Menear M, Grindrod K, Clouston K, Norton P, Legare F. Advancing knowledge translation in primary care. Can Fam Physician. 2012;58:623–7, e302-7.
11. Eccles MP, Armstrong D, Baker R, et al. An implementation research agenda. Implement Sci. 2009;4:18.

12. Centers for Disease Control and Prevention (CDC). (2007, February 16). Improving public health practice through translation research (R18). Available at: http://grants.nih.gov/grants/guide/rfa-files/rfa-cd-07-005.html. Accessed 9 May 2013.
13. Rabin BA, Brownson RC, Haire-Joshu D, Kreuter MW, Weaver NL. A glossary for dissemination and implementation research in health. J Public Health Manag Pract. 2008;14:117–23.
14. Knowledge Translation (KT) Clearinghouse. Canadian Institute of Health Research (CIHR). (2013). Available at: http://ktclearinghouse.ca/knowledgebase/glossary. Accessed 9 May 2013.
15. Smedley BD, Stith AY, Nelson AR. Unequal treatment: confronting racial and ethnic disparities in health care. Washington, DC: The National Academies Press; 2003.
16. Allison JJ. Health disparity: causes, consequences, and change. Med Care Res Rev. 2007;64:5S–6.
17. Chin MH, Walters AE, Cook SC, Huang ES. Interventions to reduce racial and ethnic disparities in health care. Med Care Res Rev. 2007;64:7S–28.
18. Kilbourne AM, Switzer G, Hyman K, Crowley-Matoka M, Fine MJ. Advancing health disparities research within the health care system: a conceptual framework. Am J Public Health. 2006;96:2113–21.
19. Smedley BD, Stith AY, Nelson AR. Unequal treatment: confronting racial and ethnic disparities in health care. Washington, DC: Institute of Medicine; 2003.
20. Hayward RA, Asch SM, Hogan MM, Hofer TP, Kerr EA. Sins of omission: getting too little medical care may be the greatest threat to patient safety. J Gen Intern Med. 2005;20:686–91.
21. Kohn LT, Corrigan JM, Donaldson MS. To err is human: building a safer health system. Washington, DC: Institute of Medicine; 1999.
22. Interagency Task Force on Antimicrobial Resistance. Co-Chairs: Centers for Disease Control and Prevention, Food and Drug Administration, National Institutes of Health. Public health action plan to combat antimicrobial resistance centers for disease control and prevention; 2011. http://www.cdc.gov/drugresistance/pdf/public-health-action-plan-combat-antimicrobial-resistance.pdf. Accessed 4 April 2014.
23. Snow V, Mottur-Pilson C, Gonzales R. Principles of appropriate antibiotic use for treatment of acute bronchitis in adults. Ann Intern Med. 2001;134:518–20.
24. Wenzel RP, Fowler 3rd AA. Clinical practice. Acute bronchitis. N Engl J Med. 2006;355:2125–30.
25. Roumie CL, Halasa NB, Grijalva CG, et al. Trends in antibiotic prescribing for adults in the United States – 1995 to 2002. J Gen Intern Med. 2005;20:697–702.
26. Metlay JP, Camargo Jr CA, MacKenzie T, et al. Cluster-randomized trial to improve antibiotic use for adults with acute respiratory infections treated in emergency departments. Ann Emerg Med. 2007;50:221–30.
27. Committee on Quality of Health Care in America. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: Institute of Medicine; 2001.
28. Berwick DM, Calkins DR, McCannon CJ, Hackbarth AD. The 100,000 lives campaign: setting a goal and a deadline for improving health care quality. JAMA. 2006;295:324–7.
29. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ. 2003;327:1459–61.
30. Auerbach AD, Landefeld CS, Shojania KG. The tension between needing to improve care and knowing how to do it. N Engl J Med. 2007;357:608–13.
31. Brennan TA, Gawande A, Thomas E, Studdert D. Accidental deaths, saved lives, and improved quality. N Engl J Med. 2005;353:1405–9.
32. Berwick D. The stories beneath. Med Care. 2007;45:1123–5.
33. Proctor EK, Powell BJ, Baumann AA, Hamilton AM, Santens RL. Writing implementation research grant proposals: ten key ingredients. Implement Sci. 2012;7:96.
34. Bhattacharyya O, Reeves S, Garfinkel S, Zwarenstein M. Designing theoretically-informed implementation interventions: fine in theory, but evidence of effectiveness in practice is needed. Implement Sci. 2006;1(Feb 23):5.

35. Effective Health Care: Getting evidence into practice. National Health Service Center for Reviews and Dissemination, Royal Society of Medicine Press. 1999;5(1). http://www.york.ac.uk/inst/crd/ehc51.pdf. Accessed Nov 2007.
36. Rogers EM. Diffusion of innovations. 5th ed. New York: Free Press; 2003.
37. Glasgow RE, Emmons KM. How can we increase translation of research into practice? Types of evidence needed. Annu Rev Public Health. 2007;28:413–33.
38. Sung NS, Crowley Jr WF, Genel M, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289:1278–87.
39. Nelson EC, Batalden PB, Huber TP, et al. Microsystems in health care: Part 1. Learning from high-performing front-line clinical units. Jt Comm J Qual Improv. 2002;28:472–93.
40. Pinsky PF, Ford M, Gamito E, et al. Enrollment of racial and ethnic minorities in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial. J Natl Med Assoc. 2008;100:291–8.
41. Ford ME, Siminoff LA, Pickelsimer E, et al. Unequal burden of disease, unequal participation in clinical trials: solutions from African American and Latino community members. Health Soc Work. 2013;38:29–38.
42. Bleich SN, Jarlenski MP, Bell CN, LaVeist TA. Health inequalities: trends, progress, and policy. Annu Rev Public Health. 2012;33:7–40.
43. The Commonwealth Fund Commission on a High Performance Health System, Why not the best? Results from the National Scorecard on U.S. Health System Performance, 2011, The Commonwealth Fund, October 2011.
44. Census Bureau. Statistical abstract of the United States: 2012: the national data book. Washington, DC: Census Bureau; 2012.
45. Zerhouni EA. Medicine. The NIH roadmap. Science. 2003;302:63–72.
46. Zerhouni EA. US biomedical research: basic, translational, and clinical sciences. JAMA. 2005;294:1352–8.
47. Chao SR. The state of quality improvement and implementation research: expert views – workshop summary. Washington, DC: The National Academies Press; 2007.
48. Shojania KG, Grimshaw JM. Evidence-based quality improvement: the state of the science. Health Aff (Millwood). 2005;24:138–50.
49. Shojania KG, McDonald KM, Wachter RM, Owens DK, editors. Closing the quality gap: a critical analysis of quality improvement strategies, Series overview and methodology, vol. 1. Rockville: Agency for Healthcare Research and Quality (US); 2004. PubMed PMID: 20734525.
50. Williams GC, Deci EL. Activating patients for smoking cessation through physician autonomy support. Med Care. 2001;39:813–23.
51. Williams GC, McGregor H, Zeldman A, Freedman ZR, Deci EL, Elder D. Promoting glycemic control through diabetes self-management: evaluating a patient activation intervention. Patient Educ Couns. 2005;56:28–34.
52. Turner RC, Waivers LE, O'Brien K. The effect of patient-carried reminder cards on the performance of health maintenance measures. Arch Intern Med. 1990;150:645–7.
53. Bodenheimer T, Lorig K, Holman H, Grumbach K. Patient self-management of chronic disease in primary care. JAMA. 2002;288:2469–75.
54. Eng E, Parker E, Harlan C. Lay health advisor intervention strategies: a continuum from natural helping to paraprofessional helping. Health Educ Behav. 1997;24:413–7.
55. Cherrington A, Ayala GX, Amick H, Scarinci I, Allison J, Corbie-Smith G. Applying the community health worker model to diabetes management: using mixed methods to assess implementation and effectiveness. J Health Care Poor Underserved. 2008;19:1044–59.
56. Swider SM. Outcome effectiveness of community health workers: an integrative literature review. Public Health Nurs. 2002;19:11–20.
57. Silka L, Cleghorn GD, Grullon M, Tellez T. Creating community-based participatory research in a diverse community: a case study. J Empir Res Hum Res Ethics. 2008;3:5–16.
58. Lawrence Mayor's Task Force. Tools for research partnerships in Lawrence, MA. Lawrence: Lawrence Mayor's Task Force; 2006. Available from: http://www.tuftsctsi.org/About-Us/CTSI-Components/Community-Engagement/~/media/B35A1D1535DB422D90E1A47544743E4E.ashx. Accessed 27 June 2013.

59. Institute of Medicine. Clinical practice guidelines: directions for a new program. Washington, DC: National Academy Press; 1990.
60. Kryworuchko J, Stacey D, Bai N, Graham ID. Twelve years of clinical practice guideline development, dissemination and evaluation in Canada (1994 to 2005). Implement Sci. 2009;4:49.
61. Grimshaw J, Eccles M, Thomas R, et al. Toward evidence-based quality improvement. Evidence (and its limitations) of the effectiveness of guideline dissemination and implementation strategies 1966–1998. J Gen Intern Med. 2006;21 Suppl 2:S14–20.
62. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA. 1999;282:1458–65.
63. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. JAMA. 2005;294:716–24.
64. Marinopoulos SS, Dorman T, Ratanawongsa N, et al. Effectiveness of continuing medical education. Evid Rep Technol Assess (Full Rep). 2007;149:1–69.
65. Davis D, O'Brien MA, Freemantle N, Wolf FM, Mazmanian P, Taylor-Vaisey A. Impact of formal continuing medical education: do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or health care outcomes? JAMA. 1999;282:867–74.
66. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. JAMA. 1995;274:700–5.
67. Forsetlund L, Bjorndal A, Rashidian A, et al. Continuing education meetings and workshops: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2009;2:CD003030.
68. Mansouri M, Lockyer J. A meta-analysis of continuing medical education effectiveness. J Contin Educ Health Prof. 2007;27:6–15.
69. Grimshaw JM, Eccles MP, Walker AE, Thomas RE. Changing physicians' behavior: what works and thoughts on getting more things to work. J Contin Educ Health Prof. 2002;22:237–43.
70. Mazmanian PE, Davis DA. Continuing medical education and the physician as a learner: guide to the evidence. JAMA. 2002;288:1057–60.
71. Centor R, Casebeer L, Klapow J. Using a combined CME course to improve physicians' skills in eliciting patient adherence. Acad Med. 1998;73:609–10.
72. Shojania KG, Silver I, Levinson W. Continuing medical education and quality improvement: a match made in heaven? Ann Intern Med. 2012;156:305–8.
73. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA. 2008;300:1181–96.
74. Fordis M, King JE, Ballantyne CM, et al. Comparison of the instructional efficacy of Internet-based CME with live interactive CME workshops: a randomized controlled trial. JAMA. 2005;294:1043–51.
75. Soumerai SB, Avorn J. Principles of educational outreach ('academic detailing') to improve clinical decision making. JAMA. 1990;263:549–56.
76. O'Brien MA, Rogers S, Jamtvedt G, et al. Educational outreach visits: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2007;4:CD000409.
77. Flodgren G, Parmelli E, Doumit G, et al. Local opinion leaders: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2011;8:CD000125.
78. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2012;6:CD000259.
79. Valente TW, Pumpuang P. Identifying opinion leaders to promote behavior change. Health Educ Behav. 2007;34:881–96.
80. Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. JAMA. 2001;285:2871–9.
81. Davidoff F, Batalden P. Toward stronger evidence on quality improvement. Draft publication guidelines: the beginning of a consensus project. Qual Saf Health Care. 2005;14:319–25.

82. Jha AK, Perlin JB, Kizer KW, Dudley RA. Effect of the transformation of the Veterans Affairs Health Care System on the quality of care. N Engl J Med. 2003;348:2218–27.
83. Centers for Medicare & Medicaid Services. About the CMS Innovation Center. Available at: http://innovation.cms.gov/About/index.html. Accessed 26 June 2013.
84. Payne TH. Computer decision support systems. Chest. 2000;118:47S–52.
85. Walton RT, Harvey E, Dovey S, Freemantle N. Computerised advice on drug dosage to improve prescribing practice. Cochrane Database Syst Rev. 2001;1:CD002894.
86. Han YY, Carcillo JA, Venkataraman ST, et al. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics. 2005;116:1506–12.
87. Nebeker JR, Hoffman JM, Weir CR, Bennett CL, Hurdle JF. High rates of adverse drug events in a highly computerized hospital. Arch Intern Med. 2005;165:1111–6.
88. Scalise D. Technology. CPOE: are you really ready? Hosp Health Netw. 2006;80:14, 6.
89. Ash JS, Sittig DF, Poon EG, Guappone K, Campbell E, Dykstra RH. The extent and importance of unintended consequences related to computerized provider order entry. J Am Med Inform Assoc. 2007;14:415–23.
90. Reckmann MH, Westbrook JI, Koh Y, Lo C, Day RO. Does computerized provider order entry reduce prescribing errors for hospital inpatients? A systematic review. J Am Med Inform Assoc. 2009;16:613–23.
91. Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. Ann Intern Med. 2008;148:111–23.
92. Werner RM, Asch DA, Polsky D. Racial profiling: the unintended consequences of coronary artery bypass graft report cards. Circulation. 2005;111:1257–63.
93. Centers for Medicare & Medicaid Services. Quality initiatives – general information. Available at: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/index.html. Accessed 26 June 2013.
94. Houle SK, McAlister FA, Jackevicius CA, Chuck AW, Tsuyuki RT. Does performance-based remuneration for individual health care practitioners affect patient care?: a systematic review. Ann Intern Med. 2012;157:889–99.
95. Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with pay-for-performance: from concept to practice. JAMA. 2005;294:1788–93.
96. Glickman SW, Ou FS, DeLong ER, et al. Pay for performance, quality of care, and outcomes in acute myocardial infarction. JAMA. 2007;297:2373–80.
97. Lindenauer PK, Remus D, Roman S, et al. Public reporting and pay for performance in hospital quality improvement. N Engl J Med. 2007;356:486–96.
98. Murray DM. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.
99. Cable G. Enhancing causal interpretations of quality improvement interventions. Qual Health Care. 2001;10:179–86.
100. Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. Qual Saf Health Care. 2003;12:47–52.
101. Gonzales R, Handley MA, Ackerman S, O'Sullivan PS. A framework for training health professionals in implementation and dissemination science. Acad Med. 2012;87:271–8.
102. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996;276:637–9.
103. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA. 2001;285:1987–91.
104. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. BMJ. 2004;328:702–8.
105. Elbourne DR, Campbell MK. Extending the CONSORT statement to cluster randomized trials: for discussion. Stat Med. 2001;20:489–96.
106. Safford MM, Shewchuk R, Qu H, et al. Reasons for not intensifying medications: differentiating "clinical inertia" from appropriate care. J Gen Intern Med. 2007;22:1648–55.

107. Foster PP, Williams JH, Estrada CA, et al. Recruitment of rural physicians in a diabetes internet intervention study: overcoming challenges and barriers. J Natl Med Assoc. 2010;102:101–7.

108. Estrada CA, Safford MM, Salanitro AH, et al. A web-based diabetes intervention for physician: a cluster-randomized effectiveness trial. Int J Qual Health Care. 2011;23:682–9.

109. Billue KL, Safford MM, Salanitro AH, et al. Medication intensification in diabetes in rural primary care: a cluster-randomised effectiveness trial. BMJ Open. 2012;2:e000959.

110. Fouad MN, Partridge E, Green BL, et al. Minority recruitment in clinical trials: a conference at Tuskegee, researchers and the community. Ann Epidemiol. 2000;10:S35–40.

111. Department of Health and Human Services National Institutes of Health. "Monitoring adherence to the NIH policy on the inclusion of women and minorities as subjects in clinical research" comprehensive report: tracking of human subjects research as reported in fiscal year 2008 and fiscal year 2009. http://orwh.od.nih.gov/research/inclusion/pdf/Inclusion-ComprehensiveReport-FY-2008-2009.pdf. Accessed 13 May 2013.

112. Flaskerud JH, Nyamathi AM. Attaining gender and ethnic diversity in health intervention research: cultural responsiveness versus resource provision. ANS Adv Nurs Sci. 2000;22:1–15.

113. LaVeist TA, Nickerson KJ, Bowie JV. Attitudes about racism, medical mistrust, and satisfaction with care among African American and white cardiac patients. Med Care Res Rev. 2000;57 Suppl 1:146–61.

114. Randall VR. Slavery, segregation and racism: trusting the health care system ain't always easy! An African American perspective on bioethics. St Louis Univ Public Law Rev. 1996;15:191–235.

115. Charatz-Litt C. A chronicle of racism: the effects of the white medical community on black health. J Natl Med Assoc. 1992;84:717–25.

116. Fouad MN. Enrollment of minorities in clinical trials: did we overcome the barriers? Contemp Clin Trials. 2009;30:103–4.

117. Wendler D, Kington R, Madans J, et al. Are racial and ethnic minorities less willing to participate in health research? PLoS Med. 2006;3:e19.

118. Casarett D, Karlawish JH, Sugarman J. Determining when quality improvement initiatives should be considered research: proposed criteria and potential implications. JAMA. 2000;283:2275–80.

119. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA. 2000;283:2701–11.

120. Lynn J, Baily MA, Bottrell M, et al. The ethics of using quality improvement methods in health care. Ann Intern Med. 2007;146:666–73.

121. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2005;2:e267.

122. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ. 2001;322:355–7.

123. Scriven M. Beyond formative and summative evaluation. In: McLaughlin MW, Phillips DC, editors. Evaluation and education: 90th yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press; 1991. p. 18–64.

124. Weston CB, McAlpine L, Bordonaro T. A model for understanding formative evaluation in instructional design. Educ Technol Res Dev. 1995;43:29–49.

125. Delbecq AL, Van de Ven AH, Gustafson DH. Group techniques for program planning: a guide to nominal group and Delphi processes. Glenview: Scott Foresman; 1975.

126. Krueger RA, Casey MA. Focus groups: a practical guide for applied research. 3rd ed. Thousand Oaks: Sage Publications; 2000.

127. Nielsen J, Mack R. Usability inspection methods. New York: Wiley; 1994.

128. Strauss A, Corbin J. Basics of qualitative research: grounded theory, procedures, and techniques. Newbury Park: Sage Publications; 1990.

129. Casebeer LL, Strasser SM, Spettell CM, et al. Designing tailored Web-based instruction to improve practicing physicians' preventive practices. J Med Internet Res. 2003;5:e20.

130. Bootzin RR. The role of expectancy in behavior change. In: White L, Turskey B, Schwartz G, editors. Placebo: theory, research, and mechanisms. New York: Guilford Press; 1985. p. 196–210.

131. Gross D. On the merits of attention-control groups. Res Nurs Health. 2005;28:93–4.
132. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health. 2011;38:65–76.
133. Creswell JW, Klassen AC, Plano Clark VL, Smith KC for the Office of Behavioral and Social Sciences Research. Best practices for mixed methods research in the health sciences. August 2011. National Institutes of Health. Available at: http://obssr.od.nih.gov/mixed_methods_research. Accessed 1 June 2013.
134. Johnson RB, Onwuegbuzie AJ, Turner LA. Toward a definition of mixed methods research. J Mix Methods Res. 2007;1:112–33.
135. Tabak RG, Khoong EC, Chambers DA, Brownson RC. Bridging research and practice: models for dissemination and implementation research. Am J Prev Med. 2012;43:337–50.
136. Brownson RC, Colditz GA, Proctor EK, editors. Dissemination and implementation research in health: translating science to practice. Oxford: Oxford Scholarship Online; Oxford University Press; 2012.
137. Reach Effectiveness Adoption Implementation Maintenance (RE-AIM). Virginia Tech University. Available at: http://www.re-aim.org/. Accessed 19 May 2013.
138. Kessler RS, Purcell EP, Glasgow RE, Klesges LM, Benkeser RM, Peek CJ. What does it mean to "employ" the RE-AIM model? Eval Health Prof. 2013;36:44–66.
139. Glasgow RE. Implementation science models (and related metrics) to help reduce health disparities. National Cancer Institute. Available at: http://cancercontrol.cancer.gov/IS/presentations/12-18-2012_Disparities%20Conference_Glasgow_508compliant.pdf. Accessed 19 May 2013.
140. Glasgow RE, Klesges LM, Dzewaltowski DA, Estabrooks PA, Vogt TM. Evaluating the impact of health promotion programs: using the RE-AIM framework to form summary measures for decision making involving complex issues. Health Educ Res. 2006;21:688–94.
141. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health. 1999;89:1322–7.
142. Glasgow RE, Magid DJ, Beck A, Ritzwoller D, Estabrooks PA. Practical clinical trials for translating research to practice: design and measurement recommendations. Med Care. 2005;43:551–7.
143. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ. 2009;180:E47–57.
144. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA. 2003;290:1624–32.
145. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis. 1967;20:637–48.
146. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Clin Epidemiol. 2009;62:499–505.
147. Selby P, Brosky G, Oh PI, Raymond V, Ranger S. How pragmatic or explanatory is the randomized, controlled trial? The application and enhancement of the PRECIS tool to the evaluation of a smoking cessation trial. BMC Med Res Methodol. 2012;12:101.
148. Green LW. What can we generalize from research on patient education and clinical health promotion to physician counseling on diet? Eur J Clin Nutr. 1999;53 Suppl 2:S9–18.
149. Green LW. The precede-proceed model of health program planning & evaluation. Available at: http://lgreen.net/index.html. Accessed 19 May 2013.
150. Bakken S, Lantigua RA, Busacca LV, Bigger JT. Barriers, enablers, and incentives for research participation: a report from the Ambulatory Care Research Network (ACRN). J Am Board Fam Med. 2009;22:436–45.
151. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review–a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy. 2005;10 Suppl 1:21–34.
152. Connelly JB. Evaluating complex public health interventions: theory, methods and scope of realist enquiry. J Eval Clin Pract. 2007;13:935–41.

153. Rycroft-Malone J, McCormack B, Hutchinson AM, et al. Realist synthesis: illustrating the method for implementation research. Implement Sci. 2012;7:33.
154. Zwarenstein M, Treweek S, Gagnier JJ, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008;337:a2390.
155. Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. BMJ. 2003;327:785–9.
156. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health. 2004;94:423–32.
157. Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials. 2000;21:167–89.
158. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet. 2002;359:781–5.
159. Ejiogu N, Norbeck JH, Mason MA, Cromwell BC, Zonderman AB, Evans MK. Recruitment and retention strategies for minority or poor clinical research participants: lessons from the Healthy Aging in Neighborhoods of Diversity across the Life Span study. Gerontologist. 2011;51 Suppl 1:S33–45.
160. Yancey AK, Ortega AN, Kumanyika SK. Effective recruitment and retention of minority research participants. Annu Rev Public Health. 2006;27:1–28.
161. Janson SL, Alioto ME, Boushey HA. Attrition and retention of ethnically diverse subjects in a multicenter randomized controlled research trial. Control Clin Trials. 2001;22:236S–43.
162. Warren-Findlow J, Prohaska TR, Freedman D. Challenges and opportunities in recruiting and retaining underrepresented populations into health promotion research. Gerontologist. 2003;43(Spec No 1):37–46.
163. Arean PA, Alvidrez J, Nery R, Estes C, Linkins K. Recruitment and retention of older minorities in mental health services research. Gerontologist. 2003;43:36–44.
164. Ashing-Giwa K, Ganz PA. Effect of timed incentives on subject participation in a study of long-term breast cancer survivors: are there ethnic differences? J Natl Med Assoc. 2000;92:528–32.
165. Gauthier MA, Clarke WP. Gaining and sustaining minority participation in longitudinal research projects. Alzheimer Dis Assoc Disord. 1999;13 Suppl 1:S29–33.
166. Parra-Medina D, D'Antonio A, Smith SM, Levin S, Kirkner G, Mayer-Davis E. Successful recruitment and retention strategies for a randomized weight management trial for people with diabetes living in rural, medically underserved counties of South Carolina: the POWER study. J Am Diet Assoc. 2004;104:70–5.
167. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986;51:1173–82.
168. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav Res Methods Instrum Comput. 2004;36:717–31.
169. Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. Am J Public Health. 2004;94:416–22.
170. Beach ML. Primer on group randomized trials. Eff Clin Pract. 2001;4:42–3.
171. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clin Trials. 2005;2:99–107.
172. Sculpher M. Evaluating the cost-effectiveness of interventions designed to increase the utilization of evidence-based guidelines. Fam Pract. 2000;17 Suppl 1:S26–31.
173. Advancing quality improvement research: challenges and opportunities – workshop summary. Institute of Medicine. The National Academies Press; 2007. www.nap.edu/catalog/11884.html. Accessed 14 July 2013.

Chapter 14
Research Methodology for Studies of Diagnostic Tests

Stephen P. Glasser

Research is what I'm doing when I don't know what I'm doing.
Wernher von Braun http://www.brainyquote.com/quotes/authors/w/wernher_von_braun.html
Prediction is very difficult, especially about the future.
http://larry.denenberg.com/predictions.html

Abstract Much of clinical research is aimed at assessing causality. However, clinical research can also address the value of new medical tests, which will ultimately be used for screening for risk factors, to diagnose a disease, or to assess prognosis. In order to be able to construct research questions and designs involving these concepts, one must have a working knowledge of this field. In other words, although traditional clinical research designs can be used to assess some of these questions, most of the studies assessing the value of diagnostic testing are more akin to descriptive observational designs, but with the twist that these designs are not aimed to assess causality, but are rather aimed at determining whether a diagnostic test will be useful in clinical practice. This chapter will introduce the various ways of assessing the accuracy of diagnostic tests, which will include discussions of sensitivity, specificity, predictive value, likelihood ratio, and receiver operator characteristic curves.

Keywords Predictive value • Sensitivity • Specificity • Receiver operator curves • Bayes' Theorem • Likelihood ratio • Net reclassification index • Test accuracy

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_14,

Introduction

Up to this point in the book, we have been discussing clinical research predominantly from the standpoint of causality. Clinical research can also address the value of new medical tests, which will ultimately be used for screening for risk factors, to diagnose a disease, or to assess prognosis. The types of research questions one might formulate for this type of research include: "How does one know how good a test is in giving you the answers that you seek?" or "What are the rules of evidence against which new tests should be judged?" In order to be able to construct research questions and designs involving these concepts, one must have a working knowledge of this field. Although traditional clinical research designs can be used to assess some of these questions, most of the studies assessing the value of diagnostic testing are more akin to descriptive observational designs, but with the twist that these designs are not aimed to assess causality, but are rather aimed at determining whether a diagnostic test will be useful in clinical practice.

Bayes Theorem

Thomas Bayes was an English theologian and mathematician who lived from 1702 to 1761. In an essay published posthumously in 1863 (by Richard Price), Bayes' offers a solution to the problem "…to find the chance of probability of its happening (a disease in the current context) should be somewhere between any two named degrees of probability" [1]. Bayes' Theorem provides a way to apply quantitative reasoning to the scientific method. That is, if a hypothesis predicts that something should occur and it does, it strengthens our belief in that hypothesis; and, conversely if it does not occur, it weakens our belief. Since most predictions involve probabilities i.e. a hypothesis predicts that an outcome has a certain % chance of occurring, this approach has also been referred to as probabilistic reasoning. Bayes' Theorem is a way of calculating the degree of belief one has about a hypothesis. Said in another way, the degree of belief in an uncertain event is conditional on a body of knowledge (this is in contrast to the traditional statistical model called the frequentist approach which does not incorporate prior knowledge in its statistical calculations). Suppose we're screening people for a disease (D) with a test that gives either a positive or a negative result (A and B, or T+ and T− respectively). Suppose further that the test is quite accurate, in the sense that, for example, it will give a positive result 95 % of the time when the disease is present (D+), i.e. $P(T+|D+) = 0.95$ (this formula asks what is the probability of the disease being present GIVEN a positive test?), or said another way, what is the probability that a person who tests positive has disease? The naive answer is 95 %; but this is wrong. What we really want to know clinically is $P(D+|T+)$, that is, what is the probability of testing positive if one has the disease; and, Bayes' theorem (or predictive value) tells us that.

In modern medicine the first useful application of Bayes' theorem was reported in 1959 [2]. Ledley and Lusted demonstrated a method to determine the likelihood that a patient had a given disease when various combinations of symptoms known to be associated with that disease were present [2]. Redwood et al. utilized Bayesian logic to reconcile seemingly discordant results of treadmill exercise testing and coronary angiography [3]. In 1977, Rifkin and Hood pioneered the routine application of Bayesian probability in the non-invasive detection of coronary artery disease (CAD) [4]. This was followed by other investigative uses of Bayesian analysis, an approach which has now become one of the common ways of evaluating all diagnostic testing.

As noted above, diagnostic data can be sought for a number of reasons beside just the presence or absence of disease. For example, the interest may be the severity of the disease, the ability to predict the clinical course of a disease, or to predict a therapy response. For a test to be clinically meaningful one has to determine how the test results will affect clinical decisions, what are its cost, risks, and what is the acceptability of the test; in other words, how much more likely will one be about this patients problem after a test has been performed than one was before the test; and, is it worth the risk and the cost? Recall, that the goal of studies of diagnostic testing seeks to determine whether a test is useful in clinical practice. To derive the latter we need to determine whether the test is reproducible, how accurate it is, whether the test affects clinical decisions, etc. One way to statistically assess test reproducibility (i.e. inter and intra-variability of test interpretation), is with a kappa statistic [5]. Note that reproducibility does not require a gold standard, while accuracy does. In order to talk intelligently about diagnostic testing, some basic definitions and understanding of some concepts is necessary.

Kappa Statistic (k)

The kappa coefficient is a statistical measure of inter-rater reliability. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Cohen's kappa measures the agreement between two raters [5].

The equation for κ is:

$$\Pr(a) - \Pr(e) / 1 - \Pr(e)$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the probability that agreement is due to chance.

If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$ (See Table 14.1). Note that Cohen's kappa measures agreement between two raters only. For a similar measure of agreement when there are more than two raters Fleiss' kappa is used [5]. An example of the use of the kappa statistic is shown in Table 14.2.

Table 14.1 Strength of agreement using the kappa statistic

Kappa	Strength of agreement
0.00	Poor
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Table 14.2 An example of the use of the kappa statistic

		Doctor A		
		No	Yes	Total
Doctor B	No	10(34.5 %)	7(24.1 %)	17(58.6 %)
	Yes	0(0.0 %)	12(41.4 %)	12(41.4 %)
Total		10(34.5 %)	19(65.5 %)	29

Kappa = (Observed agreement – Chance agreement)/(1 – Chance agreement)
Observed agreement = (10 + 12)/29 = 0.76
Chance agreement = 0.586 * 0.345 + 0.655 * 0.414 = 0.474
Kappa = (0.76 – 0.474)/(1 – 0.474) = 0.54

Definitions

Pre-test Probability

The pre-test probability (likelihood) that a disease of interest is present or not, is the index of suspicion for a diagnosis, ***before*** the test of interest is performed. This index of suspicion is influenced by the prevalence of the disease in the population of patients you are evaluating. Intuitively, one can reason that with a rare disease (low prevalence) that even with a high index of suspicion, you are more apt to be incorrect regarding the disease's presence, than if you had the same index of suspicion in a population with high disease prevalence.

Post-test Probability and Test Ascertainment

The post-test probability is one's index of suspicion ***after*** the test of interest has been performed. Let's further explore this issue as follows. If we construct a 2×2 table (Table 14.3) we can define the following variables: If disease is present and the test is positive, that test is called a true positive (TP) test (this forms the definition of test sensitivity – that is the % of TP tests in patients with the index disease). If the index disease is present and the test is negative, that is called a false negative (FN) test. Thus, patients with the index disease can have a TP or FN result (but by definition cannot have a false positive – FP, or a true negative -TN result).

Table 14.3 The relationship between disease and test result

	Abnormal test	Normal test
Disease present	True positive (TP)	False negative (FN)
Disease absent	False positive (FP)	True negative (TN)

Table 14.4 An example of the pre and post-test probability given disease prevalence and the sensitivity and specificity of a test

Pre vs post-test probability		
Prev = 10 % of 100 patients, Se = 70 %, Sp = 90 %		
	T+	**T−**
D+	7/10 (TP)	3/10 (FN)
D−	9/90 (FP)	81/90(TN)
	PV+ 7/16 = 44 % (10 % → 44 %) PV− 81/84 = 97 % (90 % → 96 %)	

Sensitivity and Specificity

The sensitivity of a test then can be written as TP/TP+FN. If the index disease is not present (i.e. it is absent) and the test is negative, this is called a true negative (TN) test (this forming the definition of specificity-that is the % of TN's in the absence of disease). The specificity of a test can then be written as TN/TN+FP. Finally, if disease is absent and the test is positive one has a false positive (FP) test. Note that the FP % is 1-specificity (that is, if the specificity is 90 % – in 100 patients without the index disease, 90 will have a negative test, which means 10 will have a positive test – i.e. FP is 10 %).

Predictive Value

Another concept is that of the predictive value (PV+ and PV−) of a test. This is asking the question differently than what sensitivity and specificity address – that is rather than asking what the TP and TN rate of a test is, the PV+ of a test result is asking how likely is it that a positive test is a true positive (TP)? i.e. TP/TP+FP (for PV− it is TN/TN+FN). See the example of the calculation of PV in Table 14.4.

Ways of Determining Test Accuracy and/or Clinical Usefulness

There are at least six ways of determining test accuracy and they are all interrelated so the determination of which to use is based on the question being asked, and one's personal preference. They are:

Sensitivity and Specificity
2×2 Tables

Predictive Value
Bayes Formula of Conditional Probability
Likelihood Ratio
Receiver Operator Characteristic Curve (ROC)

Bayes Theorem

We have already discussed sensitivity and specificity as well as the tests predictive value, and the use of 2×2 tables; and, examples will be provided at the end of this chapter. But, understanding Bayes Theorem of conditional probabilities will help provide the student interested in this area with greater understanding of the concepts involved. First let's discuss some definitions and probabilistic lingo along with some shorthand. The conditional probability that event A occurs given population B is written as P(A|B). If we continue this shorthand, sensitivity can be written as P(T+|D+) and PV+ as P(D+|T+). Bayes' Formula can be written then as follows: The post test probability of disease =

$$\frac{(\text{Sensitivity})(\text{disease prevalence})}{(\text{Sensitivity})(\text{disease prevalence})+(1-\text{specificity})(\text{disease absence})}$$

or

$$\frac{P(D+|T+)=P(T+|D+)(\text{prevalence}\,D+)}{P(T+|D+)(\text{prevalence}\,D+)P(T+|D-)P(D-)}$$

where P(D+|T+) is the probability of disease given a T+ (otherwise known as PV+), P(T+|D+) is the shorthand for sensitivity, P(T+|D–) is the FP rate or 1-specificity. Some axioms apply. For example, one can arbitrarily adjust the "cut-point" separating a positive from a negative test and thereby change the sensitivity and specificity. However, any adjustment that increases sensitivity (this then increases ones comfort that they will not "miss" any one with disease as the false negative rate necessarily falls) will decrease specificity (that is the FP rate will increase – recall 1-specificity is the FP rate). An example of this is using the degree of ST segment depression during an electrocardiographic exercise test that one has determined will identify whether the test will be called "positive" or "negative". The standard for calling the ST segment response as positive is 1 mm of ST segment depression from baseline, and in the example in Table 14.2 this yields a sensitivity of 62 % and specificity of 89 %. Note what happens when one changes the definition of what a positive test is, by using 0.5 mm ST depression as the cut-point for calling test positive or negative. Another important axiom is that the prevalence of disease in the population you are studying does not significantly influence the sensitivity or specificity of a test (to derive those variables the denominators are defined as subjects with or without the

Table 14.5 An example of calculating post test probability of disease using Bayes formula

Pre vs post-test probability

Prev = 50 % in 100 patients, Se = 70 %, Sp = 90 %

	T+	**T−**
D+	.7 × 50 = 35 (TP)	.3 × 50 = 15 (FN)
D−	.1 × 50 = 5 (FP)	.9 × 50 = 45 (TN)

PV+ 35/40 = 87 %

PV− 45/60 = 75 %

$$P(D+T+) = \frac{.7(.5)}{.7(.5)+1-.9(.5)} = \frac{.35}{.35+.05} = .87$$

Table 14.6 Estimations of pre and post test probabilities of disease given the clinical presentation

Pre vs post-test probabilities

Clinical presentation	Pre test P (%)	Post test P T+ (%)	Post test P T− (%)
Typical angina	90	98	75
Atypical angina	50	88	25
No symptoms	10	44	4

disease i.e. if you are studying a population with a 10 % disease prevalence one is determining the sensitivity of a test – against a gold standard- only in those 10 %). In contrast, PV is very dependent on disease prevalence because more individuals will have a FP test in populations with a disease prevalence of 10 % than they would if the disease prevalence was 90 %. Consider the example in Tables 14.5 and 14.6.

Receiver Operator Characteristic Curves (ROC)

The ROC is another way of expressing the relationship between sensitivity and specificity (actually 1-specificity). It plots the TP rate (sensitivity) against the FP rate over a range of "cut-point" values (actually the ROC curve is a plot of likelihood ratios – see below). It thus provides visual information on the "trade off" between sensitivity and specificity, and the area under the curve (AUC) of a ROC curve is a measure of overall test accuracy (Fig. 14.1). ROC analysis was born during WW II as a way of analyzing the accuracy of sonar detection of submarines and differentiating signals from noise [6]. In Fig. 14.2, a theoretic "hit" means a submarine was correctly identified, and a false alarm means that a noise was incorrectly identified as a submarine and so on. You should recognize this figure as the equivalent of the table above discussing false and true positives.

Another way to visualize the tradeoff of sensitivity and specificity and how ROC curves are constructed is to consider the distribution of test results in a population. In Fig. 14.3, the vertical line describes the threshold chosen for a test to be called positive or negative (in this example the right hand curve is the distribution of

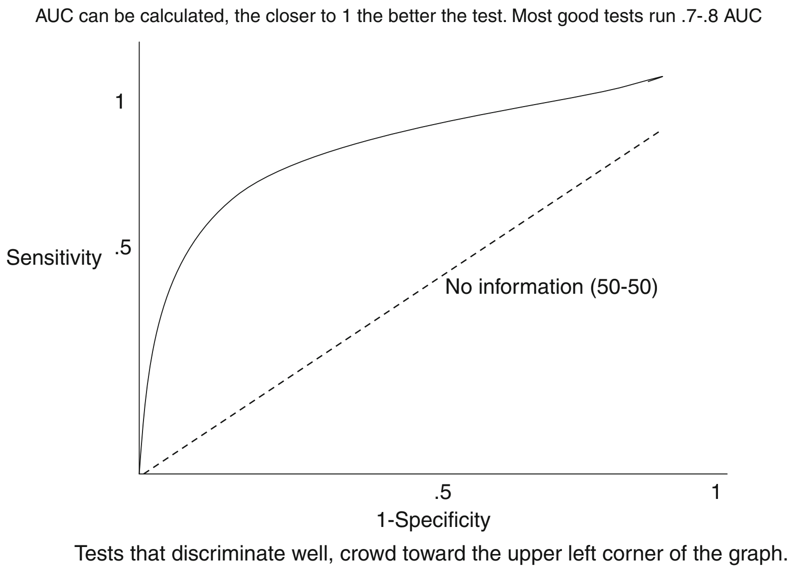
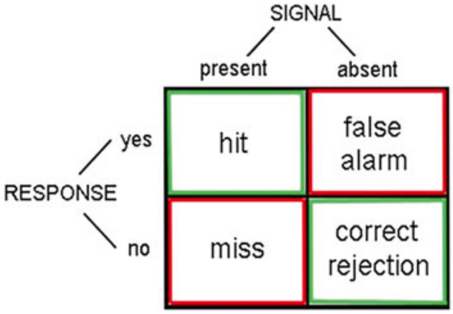


Fig. 14.1 An example of a Receiver Operator Characteristic (ROC) curve



	present	absent
yes	hit	false alarm
no	miss	correct rejection

Fig. 14.2 A diagram of the use of sonar to correctly identify submarines (http://www-psych.stanford.edu/~lera/psych115s/notes/signal/. Accessed 11/05/2013)

subjects within the population that have the disease, the left hand curve those who do not have the disease). The uppermost figure is an example of choosing a very low threshold value for separating positive from negative. By so doing, very few of the subjects with disease (recall the right hand curve) will be missed by this test (i.e. the sensitivity is high-97.5 %), but notice that 84 % of the subjects without disease will also be classified as having a positive test (false alarm or false + rate is 84 % and the specificity of the test for this threshold value is 16 %). By moving the vertical line (threshold value) we can construct different sensitivity to false + rates and construct a ROC curve as demonstrated in Fig. 14.4.

As mentioned before, ROC curves also allow for an analysis of test accuracy (a combination of TP and TN), by calculating the area under the curve as shown in the figure above. Test accuracy can also be calculated by dividing the TP and TN by all

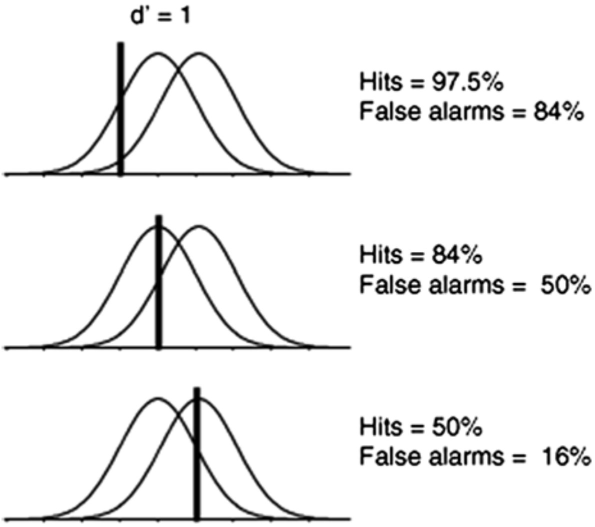


Fig. 14.3 An example of how moving the definition of positive vs negative tests alter the results of correctly identifying a target (http://www-psych.stanford.edu/~lera/psych115s/notes/signal/. Accessed 11/05/2013)

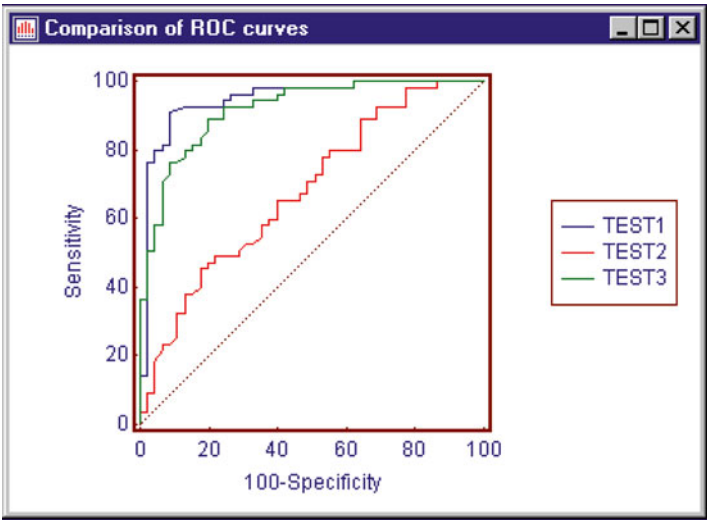


Fig. 14.4 Examples of ROC curves from three different tests (http://en.wikipedia.org/wiki/Receiver_operating_characteristic. Accessed 11/05/2013)

possible test responses (i.e. TP, TN, FP, FN). The way ROC curves can be used during the research of a new test, is to compare the new test to existent tests as demonstrated by Maisel et al. [7].

Likelihood Ratios

Positive and Negative Likelihood Ratios (PLR and NLR or LR+ and LR−) are another way of analyzing the results of diagnostic tests. Essentially, PLR is the odds that a person with a disease would have a particular test result, divided by the odds that a person without disease would have that result. In other words, how much more likely is a test result to occur in a person with disease than a person without disease. If one multiplies the pretest odds of having a disease by the PLR, one obtains the posttest odds of having that disease. The PLR for a test is calculated as the tests sensitivity/1-specificity (i.e. FP rate). So a test with a sensitivity of 70 % and a specificity of 90 % has a PLR of 7 (70/1 − 90). Unfortunately, it is made a bit more complicated by the fact that we generally want to convert odds to probabilities. That is, the PLR of 7 is really an odds of 7 to 1 and that is more difficult to interpret than a probability (the probability from a 7:1 odds is 87.5 %, see below). Recall that odds of an event are calculated as the number of events occurring, divided by the number of an events *not* occurring (i.e. non events, or p/p − 1). So if blood type O occurs in 42 % of people, the odds of someone having a blood type of O are .42/1 − .42 i.e. the odds of a randomly chosen person having blood type O is .72:1. Probability is calculated as the odds/odds + 1, so in the example above .72/1.72 = 42 % (or .42 – that is one can say the odds have having blood type O is .72 to 1 or the probability is 42 %-the latter is easier to understand for most). Recall, that probability is the extent to which something is likely to happen. To review, take an event that has a 4 in 5 probability of occurring (i.e. 80 % or .8). The odds of its occurring is 0.8/1 − 0.8 or 4:1. Odds then, are a ratio of probabilities. Note that an odds ratio (often used in the analysis of clinical trials) is also a ratio of odds.

To review:

The likelihood ratio of a positive test (LR+) is usually expressed as

$$\text{Sensitivity} / 1 - \text{Specificity}$$

and the LR− is ***usually*** expressed as

$$1 - \text{Sensitivity} / \text{Specificity}$$

If one has estimated a pretest odds of disease, one can multiply that odds by the LR to obtain the post test odds, i.e.:

$$\text{Post - test odds} = \text{pre - test odds} \times \text{LR}$$

To use an exercise test example consider the sensitivity for the presence of CAD (by coronary angiography) based on 1 mm ST segment depression. In this aforementioned example, let's assume that the sensitivity of a "positive" test is 70 % and the specificity is 90 % (PLR = 7; NLR = .33). Let's assume that based upon our

Table 14.7 Different ways of calculating Likelihood Ration (LR)

End point	LR	Ratio	Se:Sp
D+ for T+	**LR+**	**%D+ with T+** **%D− with T+**	**Se/1−SP** **TP/FP**
D− for T−	LR−	%D− with T− %D+ with T−	Sp/1−Se TN/FN
D− for T+	1/LR+	%D− with T+ %D+ with T+	1−Sp/Se FP/TP
D+ for T−	**1/LR−**	**%D+ with T−** **%D− with T−**	**1−Se/Sp** **FN/TN**

history and physical exam we feel the chance of a patient having CAD before the exercise test is 80 % (0.8). If the exercise test demonstrated 1 mm ST segment depression, your post-test odds of CAD would be .8 × 7 or 5.6 (to 1). The probability of that patient having CAD is then 5.6/1 + 5.6 = .85 (85 %). Conversely if the exercise test did not demonstrate 1 mm ST segment depression the odds that the patient did not have CAD is .33 × 7 = 2.3 (to 1) and the probability of his not having CAD is 70 %. In other words *before* the exercise test there was an 80 % chance of CAD, while *after* a positive test it was 85 %. Likewise before the test, the chance of the patient *not* having CAD was 20 %, and if the test was negative it was 70 %.

To add a bit to the confusion about using LRs, there are two lesser-used derivations of the LR as shown in Table 14.7. One can usually assume that if not otherwise designated, the descriptions for PLR and NLR above apply. But, if one wanted to express the results of a negative test in terms of the chance that the patient **has** CAD (despite a negative test) rather than the chance that he **does not** have disease given a negative test; or wanted to match the NLR with NPV (i.e. the likelihood that the patient does NOT have the disease given a negative test result) an alternative definition of NLR can be used (of course one could just as easily subtract 70 % form 100 % to get that answer as well). To make things easier, a nomogram can be used instead of having to do the calculations [8].

In summary, the usefulness of diagnostic data depends on making an accurate diagnosis based upon the use of diagnostic tests, whether the tests are radiologic, laboratory based, or physiologic. The questions to be considered by this approach include: "How does one know how good a test is in giving you the answers that you seek?", and "What are the rules of evidence against which new tests should be judged?" Diagnostic data can be sought for a number of reasons including: diagnosis, disease severity, to predict the clinical course of a disease, to predict therapy response. That is, what is the probability my patient has disease x, what do my history, physical exam, and baseline laboratory data tell me, what is my threshold for action, and how much will the available tests help me in patient management. An example of the use of diagnostic research is provided by Miller and Shaw, which demonstrates how the coronary artery calcium (CAC) score can be stratified by age and the use of the various definitions described above [9].

Beyond the ROC Curve

Over 30 years after the construction of the first multivariable risk prediction model predicting the probability of developing cardiovascular disease (CVD) new risk factors that can predict CVD and that can be incorporated into risk assessment algorithms has progressed. An individual's age, baseline levels of systolic and diastolic blood pressure and serum cholesterol, smoking and diabetes status are all useful predictors of the CVD risk over a reasonable future time period, typically 1–10 years. Quantification of vascular risk is accomplished through risk equations or risk score sheets that have been developed on the basis of observations from large cohort studies. For example, the Framingham risk score has been routinely applied, validated and calibrated for use. However, CVD risk prediction is an ongoing work in progress and new risk factors or markers are being identified and proposed constantly. The critical question arises is to how to evaluate the usefulness of a new marker? Four initial decisions that guide the process are:

- defining the population of interest
- defining the outcome of interest
- choosing how to incorporate the competing pre-existing set of risk factors
- selecting the appropriate model and tests to evaluate the incremental yield of a new biomarker

Since, none of the numerous new markers proposed comes close in magnitude to the necessary levels of association, some have argued that we need to wait for new and better markers; others have sought model performance measures beyond the AUC calculated from a ROC curve to evaluate the usefulness of markers. For example, the Net Reclassification Index (or Improvement-NRI), focuses on reclassification tables constructed separately for participants with and without events, and quantifies the correct movement in categories – upwards for events and downwards for non-events. In its simplest terms, the NRI is defined as a measure of the net % of those who do or do not develop an endpoint within a given time period that are correctly reclassified to a different category when a new risk factor is added to the risk estimation [1]. Again in its simplest terms, one can construct a 2×2 table and assess an endpoint, then add a new risk factor and reassess. The % improvement in TP and TN is the NRI. One example of this is the use of the coronary artery calcium (CAC) score to reclassify the patients risk say from that predicted by the FRS. The addition of a CAC score in one study, altered conventional risk determination (Framingham Risk Score [FRS]) such that the posttest probability could reclassify a patient to a new category of risk.

Although the data using the NRI are conceptually appealing for patient care, there are still many unanswered questions with substantial clinical implications that will need to be addressed prior to using this reclassification in clinical practice.

Screening Testing

Screening tests are ubiquitous in contemporary practice, yet the principles of screening are widely misunderstood. Screening is the testing of apparently well people to find those at increased risk of having a disease or disorder. Those identified are sometimes then offered a subsequent diagnostic test or procedure, or, in some instances, a treatment or preventive medication. Looking for additional illnesses in those with medical problems is termed case finding. Although an earlier diagnosis generally has intuitive appeal, earlier might not always be better, or worth the cost. For tests with continuous variables – e.g., blood glucose – sensitivity and specificity as mentioned prior, are inversely related; where the cutoff for abnormal is placed should indicate the clinical effect of wrong results. As also prior mentioned, the prevalence of disease in a population affects screening test performance: in low-prevalence settings, even very good tests have poor positive predictive values. Hence, knowledge of the approximate prevalence of the index disease is a prerequisite to interpreting screening test results.

Screening differs from the traditional clinical use of tests in several important ways. Ordinarily, patients consult with clinicians about complaints or problems; and, this prompts testing to confirm or exclude a diagnosis. Because the patient is in pain and requests help, the risk and expense of tests are usually deemed acceptable by the patient. By contrast, screening engages apparently healthy individuals who are not seeking medical help (and who might prefer to be left alone). Hence, the cost, injury, and stigmatization related to screening are especially important (though often ignored in our zeal for earlier diagnosis). Furthermore, the medical and ethical standards of screening should be, correspondingly, higher than with diagnostic tests. Bluntly put: every adverse outcome of screening is iatrogenic and entirely preventable; thus, screening has a darker side that is often overlooked.

Guidelines for Publishing or Assessing Research in Diagnostic Tests

Finally, just as there are guidelines for publishing and assessing published articles addressing clinical and observational trials (see Chaps. 3 and 19) there are also guidelines for publishing studies of new diagnostic tests. McReid et al. have suggested seven methodological standards for diagnostic tests [2] as follows.

- *Spectrum Composition*: i.e. if one changes the population under study one can change the tests diagnosticity, thus in assessing the results of a new diagnostic test, information on age and sex distribution, presenting symptoms and/or disease stage, and eligibility criteria for study patients should be included in published works.
- *Pertinent Subgroups*: Se and Sp represent average values for a population. Unless the condition is narrowly defined, the indices may vary for different medical subgroups, thus these subgroups should be clearly described.

- *Avoidance of Workup Bias*: patients with a positive or negative "gold standard" diagnostic tests might be preferentially referred to evaluate the diagnosticity of a newly reported test. For example, a new DNA test to detect the breast cancer gene was administered to biopsy proven breast cancer and cancer-free controls. Since the biopsy may be ordered preferentially in women with a family history of breast cancer, the cases selected for the new test will be enriched by a clinical factor that itself may be associated with the new DNA test.
- *Avoidance of Review Bias*: The new test needs to be interpreted independently of other tests, and the new test and the gold standard test need to be interpreted separately by persons unaware of the results of the other (akin blinding in clinical trials).
- *Precision of Results for Test Accuracy*: Like any other research, point estimates should have confidence limits reported.
- *Presentation Of Indeterminate Results*: Not all tests come out Yes or No. Sometimes they are equivocal or indeterminate. The frequency of these results may limit the tests applicability, or make it cost more because additional test are then needed. Finally,
- *Test Reproducibility*: must be reported.

References

1. Bayes T. An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond B Biol Sci. 1763;53:370–418.
2. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science. 1959;130:9–21.
3. Redwood DR, Borer JS, Epstein SE. Whither the ST segment during exercise. Circulation. 1976;54:703–6.
4. Rifkin RD, Hood Jr WB. Bayesian analysis of electrocardiographic exercise stress testing. N Engl J Med. 1977;297:681–6.
5. McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, For GG, et al. Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). CMAJ. 2004;171:1369–73. PMC527344.
6. Green DM, Swets JM. Signal detection theory and psychophysics. New York: Wiley; 1966.
7. Maisel AS, Krishnaswamy P, Nowak RM, McCord J, Hollander JE, Duc P, Omland T, Storrow AB, Abraham WT, Wu AH, Clopton P, Steg PG, Westheim A, Knudsen CW, Perez A, Kazanegra R, Herrmann HC, McCullough PA, Breathing Not Properly Multinational Study, I. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. N Engl J Med. 2002;347:161–7. doi:10.1056/NEJMoa020233.
8. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med. 1975;293:257. doi:10.1056/NEJM197507312930513.
9. Miller DD, Shaw LJ. Coronary artery disease: diagnostic and prognostic models for reducing patient risk. J Cardiovasc Nurs. 2006;21:S2–16; quiz S17–9.

Chapter 15
Statistical Power and Sample Size: Some Fundamentals for Clinician Researchers

J. Michael Oakes

Surgeon: Say, I've done this study but my results
are disappointing.
Statistician: How so?
Surgeon: The p-value for my main effect was 0.06.
Statistician: And?
Surgeon: I need something less than 0.05 to get tenure.

Abstract This chapter aims to arm clinical researchers with the necessary conceptual and practical tools (1) to understand what sample size or power analysis is, (2) to conduct such analyses for basic low-risk studies, and (3) to recognize when it is necessary to seek expert advice and input. I hope it is obvious that this chapter aims to serve as a general guide to the issues; specific details and mathematical presentations may be found in the cited literature. Additionally, it should be obvious that this discussion of statistical power is focused, appropriately, on quantitative investigations into real or hypothetical effects of treatments or interventions. It does not address *qualitative* study designs. The ultimate goal here is to help practicing clinical researcher *get started* with power analyses.

Keywords Statistical power • Inference • Standard error • Type I and II error • Minimal detectable effect

J.M. Oakes, Ph.D. (✉)
Division of Epidemiology and Community Health, School of Public Health,
University of Minnesota, Minneapolis, MN, USA
e-mail: oakes007@umn.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_15,

Introduction

My experience as both and educator and collaborator is that clinical researchers are frequently perplexed if not unnerved by questions of statistical power, detectable effect, number-needed-to-treat, sample size calculations, and related concepts. Those who have taken a masters-level biostatistics course may even become paralyzed by authoritative cautions, supporting the quip that a little knowledge can be a dangerous thing. Unfortunately, anxiety and misunderstanding seem to push some to ignore the issues while others appear rigid in their interpretations, rejecting all 'under-powered' studies as useless. Neither approach is helpful to researchers or medical science.

I do not believe clinician researchers, especially, are to blame for the trouble. My take is that when it comes to statistical power and related issues, instructors, usually biostatisticians, are too quick to present equations and related algebra instead of the underlying concepts of uncertainty and inference. Such presentations are understandable since the statistically-minded often *think* in terms of equations and are obviously equipped with sufficient background information and practice to make sense of them. But the same is not usually true of clinicians or perhaps even some epidemiologists. Blackboards filled with Greek letters and algebraic expressions, to say nothing of terms like 'sampling distribution,' only seem to intimidate if not turn-off students eager to understand and implement the ideas. What is more, I have come across strikingly few texts or articles aimed at helping clinician-researchers understand key issues. Most seem to address only experimental (e.g., drug trial) research, offer frightening cautions, or consider only painfully simple studies. Little attention is paid to less glorious but common clinical studies such as sample-survey research or perhaps the effects of practice/cultural changes to an entire clinic. Too little has written about the conceptual foundations of statistical power, and even less of this is tailored for clinician-researchers.

I find that clinical researchers gain a more useful understanding of, and appreciation for, the concepts of statistical power when the ideas are first presented with some utilitarian end in mind, and when the ideas are located in the landscape of inference and research design. Details and special-cases are important, but an emphasis must be placed on simple and concrete examples relevant to the audience. Mathematical nuance and deep philosophical issues are best reserved for the few who express interest. Still, I agree with Baussel and Li [1] who write,

> … a priori consideration of power is so integral to the entire design process that its consideration should not be delegated to individuals not integrally involved in the conduct of an investigation…

Importantly, emphasis on concepts and understanding may also be sufficient for clinical researchers since I believe the following three points are critical to a successful power-analysis:

1. *The More, the Merrier* – Except for exceptional cases when study subjects are exposed to more than minimal risk, there is hardly any *pragmatic* argument for

not enrolling as many subjects as the budget permits. Over-powered studies are not much of a threat, especially when authors and readers appreciate the abundant limitations of p-values and other summary measures of 'significance.' While perhaps alarming, I have found analytic interest in subgroup comparisons or other 'secondary' aims to be universal; few researchers are satisfied when 'real' analyses are limited to main study hypotheses. It follows that more subjects are always needed. But let me be clear: when risk is elevated, clinical researchers must seek expert advice.

2. *Use Existing Software* – Novice study designers should rely on one or more of the high-quality and user-friendly software packages available for calculating statistical power. Novice researchers should not attempt to derive new equations nor should they attempt to implement any such equation into a spreadsheet package. The possibility of error is too great and efforts to 're-invent the wheel' will likely lead to mistakes. Existing software packages have been tested and will give the correct answer, provided researchers input the correct information. This means, of course, that the researcher must understand the function of each input parameter and the reasonableness of the values entered.
3. *If No Software, Seek Expert* – If existing sample-size software cannot accommodate a particular study design or an analysis plan, novice researchers should seek expert guidance from biostatistical colleagues or like-minded scholars. Since existing software accommodates many (sophisticated) analyses, exceptions mean something unusual must be considered. Expert training, experience, and perhaps an ability to simulate data are necessary in such circumstances. Expert advice is also necessary when risks of research extend beyond the minimal threshold.

The upshot is that clinical researchers need to minimally know what sort of sample size calculation they need and, at most, what related information should be entered into existing software. Legitimate and accurate *interpretation* of output is then paramount, as it should be. Concepts matter most here, and are what seem to be retained anyway [2].

Accordingly, this chapter aims to arm clinical researchers with the necessary conceptual and practical tools (1) to understand what sample size or power analysis is, (2) to conduct such analyses for basic low-risk studies, and (3) to recognize when it is necessary to seek expert advice and input. I hope it is obvious that this chapter aims to serve as a general guide to the issues; specific details and mathematical presentations may be found in the cited literature. Additionally, it should be obvious that this discussion of statistical power is focused, appropriately, on quantitative investigations into real or hypothetical effects of treatments or interventions. I do not address *qualitative* study designs. The ultimate goal here is to help practicing clinical researcher *get started* with power analyses. Alternative approaches to inference and 'statistical power' continue to evolve and merit careful consideration if not adoption, but such a discussion is far beyond the simple goals here; see [3, 4].

Fundamental Concepts

Inference

Confusion about statistical power often begins with a misunderstanding about the point of conducting research. In order to appreciate the issues involved in a power calculation, one must appreciate that the goal of research is to draw credible inferences about a phenomena under study. Of course, drawing credible inferences is difficult because of the many errors and complications that can cloud or confuse our understanding. Note that, ultimately, power calculations aim to clarify and quantify some of these potential errors.

To make issues concrete, consider patient A with systolic pressure of 140 mmHg and, patient B, with a reading of 120 mmHg. Obviously, the difference between these two readings is 20 mmHg. Let us refer to this difference as '*d*'. To sum up, we have

$$140 - 120 = d$$

Now, as sure as one plus one equals two, the measured difference between the two patient's BPs is 20. Make no mistake about it, the difference is 20, not more, not less.

So, what is the issue? Well, as any clinician knows either or both the blood-pressure measures could (probably do!) incorporate error. Perhaps the cuff was incorrectly applied or the clinician misread the sphygmomanometer. Or perhaps the patient suffers white-coat hypertension making the office-visit measure different from the patient's 'true' measure. Any number of measurement errors can be at work making the calculation of the observed difference between patients an error-prone measure of the true difference, symbolized by Δ, the uppercase Greek-letter 'D', for True or philosophically perfect difference.

It follows that what we actually measure is a mere *estimate* of the thing we are trying to measure, the True or parameter value. We measure blood-pressures in both patients and calculate a difference, 20, but no clinician will believe that the true or real difference in pressures between these two individuals is precisely 20 now or for all time. Instead, most would agree that the quantity 20 is an estimate of the true difference, which we may believe is 20, plus or minus 5 mmHg, or whatever. And that this difference changes over time if not place.

This point about the observed difference of 20 being an estimate for the true difference is key. One takes measures, but appreciates that imprecision is the rule. How can we gauge the degree of measurement error in our estimate of $d = 20 \rightarrow \Delta$?

One way is to take each patient's blood-pressures (BP) multiple times and, say, average them. It may turn out that patient A's BP was measured as 140, 132, 151, 141 mmHg, and patient B might have measures 120, 121, 123, 119, 117. The average of patient A's four measurements is, obviously, 141 mmHg, while patient B's five measurements yield an average of 120 mmHg. If we use these presumably more accurate average BPs, we now have this

$$140 - 120 = 21 = d*$$

where d^* is used to show that this ‘d’ is based on a different calculation (e.g., averages) than the previously discussed ‘d’.

How many estimates of the true difference do we need to be comfortable making claims about it? Note that the p-value from the appropriate t-test is less than 0.001. What does this mean? Should we take more measures? How accurate do we need the difference in blood pressure to be before we are satisfied that patient A’s BP is higher than patient B’s? Should we worry that patient A’s BP was much more variable (standard deviation = 7.8) than patient B’s (standard deviation = 2.2)? If patient A is male and patient B female, can we generalize and say that, on average, males have high BP than females? If we are a little wrong about the differences in blood pressures, which is more important: claiming there is no difference when in fact there is one, or claiming there is a difference when in fact there is not one? It is questions like these that motivate our discussion of statistical power.

The basic goal of a ‘power analysis’ is to appreciate *approximately* how many subjects are needed to detect a *meaningful* difference between two or more experimental groups. In other words, the goal of power analysis is to consider natural occurring variance of the outcome variable, errors in measurement, and the impact of making certain kinds of inferential errors (e.g., claiming a difference when in truth the two persons or groups are identical). Statistical power calculations are about inference, or making (scientific) leaps of faith from real-world observations to statements about the underlying truth.

Notice above, that I wrote ‘approximately.’ This is neither a mistake nor a subtle nuance. Power calculations are useful to determine if a study needs 50 or 100 subjects; the calculations are not useful in determining whether a study needs 50 or 52 subjects. The reason is that power calculations are loaded with assumptions, too often hidden, about distributions, measurement error, statistical relationships and perfectly executed study designs. As mentioned above, it is rare for such perfection to exist in the real world. Believing a given power analysis is capable of differentiating the utility of a proposed study within a degree of a handful of study subjects is an exercise in denial and is sure to inhibit scientific progress.

I also wrote that power was concerned with differences between ‘two groups.’ Of course study designs with more groups are possible and perhaps even desirable. But power calculations are best done by keeping comparisons simple, as when only two groups are involved. Furthermore, this discussion centers on elementary principles and so simplicity is paramount.

The other important word is ‘meaningful’. It must be understood that power calculations offer nothing by way of *meaning*; manipulation of arbitrary quantities through some algebraic exercise is a meaningless activity. The meaningfulness of a given power calculation can only come from scientific/clinical expertise. To be concrete, while some may believe a difference of, say, 3 mmHg of systolic blood pressure between groups is important enough to act on, others may say such a difference is not meaningful *even if* it is an accurate measure of difference. The proper attribution of meaningfulness, or perhaps importance or utility, requires extra-statistical knowledge. Clinical expertise is paramount.

Estimator	Standard Error
Sample mean	$\sqrt{\sigma^2/n}$
Difference between independent sample means	$\sqrt{\sigma^2\left(1/n_1 + 1/n_2\right)}$
Binomial proportion	$\sqrt{p(1-p)/n}$
Log Odds-ratio	$\sqrt{1/a + 1/b + 1/c + 1/d}$
Difference between two means in a Group-randomized trial	$\sqrt{\dfrac{2\left[\sigma^2/m + \tau^2\right]}{g}}$

Fig. 15.1 Common standard error formulas

Standard Errors

A fundamental component of statistical inference is the idea of 'standard error.' As an *idea*, a standard error can be thought of as the standard deviation of a test statistic in the sampling distribution. You may be asking, what does this mean?

Essentially, our simplified approach to inference is one of replicating a given study over and over again. This replication is not actually done, but is instead a thought, experiment, or theory that motivates inference. The key is to appreciate that for each hypothetical and otherwise identical study we observe a treatment effect or some other outcome measure. Because of natural variation and such, for some studies the test statistic is small/low, for others, large/high. Hypothetically, the test statistic is distributed in a bell-shaped curve, with one point/measure for each hypothetical study. This distribution is called the *sampling distribution*. The standard deviation (or spread) of this sampling distribution is the standard error of the test statistic. The smaller the standard deviation, the smaller the standard error.

We *calculate* standard errors in several ways depending on the study design and the chosen test statistics. Standard error formulas for common analytic estimators (i.e., tests) are shown in Fig. 15.1. Notice the key elements of each standard error formula are the variance of the outcome measure, σ^2, and sample size, n. Researchers must have a sound estimate of the outcome measure variance at planning. Reliance on existing literature and expertise is a must. Alternative approaches are discussed by Browne [5].

Since smaller standard errors are usually preferred (as they imply a more precise test statistic), one is encouraged to use quality measurement tools and/or larger sample sizes.

Hypotheses

A fundamental idea is that of the 'hypothesis' or 'testable conjecture.' The term 'hypothesis' may be used synonymously with 'theory'. A necessary idea here is that the researcher has a reasoned and *a priori* guess or conjecture about the outcome of their analysis or experiment. The *a priori* (or in advance) aspect is critical since power is done in the planning stage of a study.

For purposes here, hypotheses may be of just two types: the null and the alternative. The null hypothesis is, oddly, what is *not expected* from the study. The alternative hypothesis is what is expected given one's theory. This odd reversal of terms or logic may be a little tricky at first but everyone gets used to it. Regardless, the key idea is that researchers marshal information and evidence from their study to either confirm or disconfirm (essentially reject) their *a priori* null hypothesis. For us, a study is planned to test a theory by setting forth a null and alternative hypothesis and evaluating data/results accordingly. Researchers will generally be glad to observe outcomes that refute null hypotheses.

Several broad kinds of hypotheses are important for clinical researchers but two merit special attention:

1. Equality of groups – The null hypothesis is that the, say, mean in the treatment group is strictly equal to the mean in the control group; symbolically $\mu_T=\mu_C$, where μ_T represents the mean of the treatment group and μ_C represents the mean of the control group. The analysis conducted aims to see if the treatment is strictly different from control; symbolically $\mu_T\neq\mu_C$. As can be imagined, this strict equality or difference hypothesis is not much use in the real world.
2. Equivalence of groups – In contrast to the equality designs, equivalence designs do not consider just any difference to be important, even if statistically significant! Instead, equivalence studies require that the identified difference be clinically meaningful, above some pre-defined value, d. The null hypothesis in equivalence studies is that the (absolute value of) the difference between treatment and control groups be larger than some meaningful value; symbolically, $|\mu_T-\mu_C|\geq d$. The alternative hypothesis is then that the observed difference is smaller than the predefined threshold value d, or in symbols $|\mu_T-\mu_C|<d$. If the observed is less than d, then two 'treatments' are viewed as equivalent, though this does not mean strictly equal.

Finally, it is worth pointing out that authors typically abbreviate the term null hypothesis with H_0 and the alternative hypothesis with H_A.

	Mother Nature or True State of Null Hypothesis	
Researcher's Inference	H_0 is True	H_0 is False
Reject H_0	***Type I error*** *probability* = α	Correct Inference *probability* = $1 - \beta$ *Power* (H_A)
Accept H_0	Correct Inference *probability* = $1 - \alpha$	***Type II error*** *probability* = β

Fig. 15.2 Type I and Type II errors

Type I and Type II Error

When it comes to elementary inference, it is useful to define two kinds of errors. Using loose terms, we may call them errors of commission and omission, with respect to stated hypotheses.

Errors of commission are those of inferring a relationship between study variables when in fact there is not one. In other words, errors of commission are rejecting a null hypothesis (no relationship) when in fact it should have been accepted it. In other words, you have done something you should not have.

Errors of omission are those of not inferring a relationship between study variables when in fact there is a relationship. In other words, not rejecting a null in favor of the alternative, when in fact the alternative (a relationship) was correct. That is, you have failed to do something you should have.

The former – errors of commission – are called Type I errors. The latter, Type II errors. A simple figure is useful for understanding their inter-relationship, as shown in Fig. 15.2. Statistical researchers label Type I error α, the Greek letter 'a' or alpha. Type II errors are labeled β, the Greek letter 'b' or beta (the first and second letters of the Greek alphabet).

Both Type I and Type II errors are quantified as probabilities. The probability of incorrectly rejecting a true null hypothesis – or accepting that there is a relationship when in fact there is not – is α (ranging from 0 to 1). So, Type I error may be 0.01, 0.05 or any other such value. The same goes for Type II error.

For better or worse, by convention researchers typically plan studies with an Type I error rate of 0.05, or 5 %, and a Type II error rate of 0.20 (20 %) or less. Notice this implies that making an error of commission (5 % alpha or Type I error) is four times *more* worrisome than making an error of omission (20 % beta or Type II error). By convention, we tolerate less Type I error than Type II error. Essentially, this relationship reflects the conservative stance of science: scientists should accept the null (no relationship) unless there is strong evidence to reject it and accept the alternative hypothesis. That is the scientific method.

Statistical Power

We can now define statistical power. Technically, power is the complement of the Type II error (i.e., the difference between 1 and the amount of Type II error in the study). A simple definitional equation is,

$$\text{Power} = 1 - \beta.$$

Statistical power is, therefore, about the probability of correctly rejecting a null hypothesis when in fact one should do so. It is a population parameter, loosely explained as a study's ability or strength to reject the null when doing so is appropriate. In other words, power is about a study's ability to find a relationship between study variables (e.g., treatment effect on mortality) when in fact there is such a relationship. Note that power is a function of the alternative hypothesis; which essentially means that the larger the (treatment) effect, the more power to detect it. It follows that having more power is usually preferred since researchers want to discover new relationships between study variables. Insufficient power means some existing relationships go undetected. This is why underpowered studies are so controversial; one cannot tell if there is in fact no relationship between two study variables or whether the study was not sufficiently powered to detect the relationship; inconclusive studies are obviously less than desirable.

Given the conventional error rates mentioned above (5 % Type I and 20 % Type II) we can now see where and why the conventional threshold of 80 % power for a study obtains: it is simply

$$\text{Power} = 1 - \beta = 1 - 0.20 = 0.80$$

To be clear, 80 % statistical power means that if everything in the study goes as planned and the alternative hypothesis in fact is true, there is an 80 % chance of observing a statistically significant result and a 20 % chance of erroneously missing it. All else equal, lower Type II error rates mean more statistical power.

Power and Sample Size Formula

There are a large number of formulae and approaches to calculating statistical power and related concepts, and many of these are quite sophisticated. It seems useful however to write down a/the very basic formula and comment on it. Such foundational ideas serve as building blocks for more advanced work. The basic power formula may be written as,

$$Z_{1-\alpha/2} + Z_{Power} = \frac{\Delta}{SE(\Delta)}$$

where $Z_{\alpha/2}$ is the value of Z for a given $\alpha/2$ Type I error rate, Z_{Power} is the value of Z for a given power value (i.e., 1 – Type II error rate), Δ is the minimal detectable effect for some outcome variable (discussed below), and *SE*(Δ) is the standard error for the same outcome variable.

Let us now explore each of the four (just four!) basic elements in more detail. In short, the equation states that the (transformed) probability of making the correct inference equals the effect of some intervention divided by the appropriate standard error.

The term $Z_{\alpha/2}$ is the value of a Z statistic (often found in the back of basic statistics textbooks) for the Type I error rate divided by two, for a two-sided hypothesis test. If Type I error rate is 0.05, the value of this element is 0.975. Looking up the value of Z shows that the Z at 0.975 is 1.96.

The term Z_{Power} is the value of the Z statistic, a specified level of power. Type II error is often set a 20 % (or 0.20), which yields a Z_{Power} of 0.84.

We may now rewrite the equation for use when two-sided Type I error is 5 % and power is set at 80 % (Type II error is 20 %),

$$1.96 + 0.84 = \frac{\Delta}{SE(\Delta)}$$

The other two elements in the equation above depend on the data and/or theory. The critical part is the standard error of the outcome measure, symbolized as *SE*(Δ). This quantity depends on the study design and the variability of the outcome measure under investigation. If may be helpful to regard this quantity as the noise that is recorded in the outcome measure. Less noise means a more precise outcome measure; and, the more precision the better.

It should now be easy to see that the key part of the formula is the standard error, and thus two elements really drive statistical power calculations: variance of the outcome measure, σ^2, and sample size, *n*. The rest is more or less given, although the importance of the study design and statistical test cannot be over emphasized. It follows that for any given design researchers should aim to decrease variance and increase sample size. Doing either or both reduces the minimal detectable effect, Δ, which is generally a good thing.

Minimal Detectable Effect

As mentioned above, applied or collaborating statisticians rarely directly calculate the statistical power of a given study design. Instead, we typically ask clinician researchers how many subjects can be recruited given budget constraints and then using the conventional thresholds of 80 % power and 5 % Type I error rates calculate the study's minimum detectable difference [6]. In other words, given that (1) most find 80 % power and 5 % Type I error satisfactory and (2) that budgets are always tight, there is no point in calculating power or how many subjects are needed. Instead the values of 80 %, 5 %, and number of subject's affordable, along with the variance and other information are taken as given or immutable. The formula is algebraically manipulated to yield the smallest or minimal study effect (on the scale of the outcome measure) that is to be expected.

$$\Delta = SE(\Delta)\left[Z_{1-\alpha/2} + Z_{Power} \right]$$

For the conventional Type I and II error rates, the formula is simply

$$\Delta = SE(\Delta) * 2.8.$$

If this value is clinically meaningful – that is, not as large as to be useless – then the study is well-designed. Notice, one essentially substitutes any appropriate standard error. Again, standard errors are a function of study design (cross-sectional, cohort, or experiment study, etc.) It is worth noting that there are some subtle but important aspects to this approach; advanced learners may begin with the insights of Greenland [7].

P-Values and Confidence Interval

P-values and confidence intervals are practically related and convey a sense of uncertainty about an effect estimate. There remains a substantial degree of controversy about the utility or misuse of p-values as a measure of meaning [8–10], but the key idea is that some test statistic, perhaps Z or t, which is often the ratio of some effect estimate divided by its standard error, is assessed against a threshold value in a Z-table, say Z of 0.05 which is 1.96. If the ratio of the effect estimate divided by its standard error is greater than 1.96 (which is 1.96 standard deviations away from mean of the sample distribution) then we say the estimated effect is unlikely to arise by chance if the null hypothesis were in fact true… that is, the estimated effect is statistically significant.

Confidence intervals, often called 95 % confidence intervals, are another measure of uncertainty about estimated effects [11]. Confidence intervals are often written as the estimated mean or other statistic of the effect plus or minus some amount, such as 24 ± 11, which is to say the lower 95 % confidence interval is $24 - 11 = 13$ and the upper 95 % confidence interval is $24 + 11 = 35$. In other words, in 95 out of 100 replications of the study being conducted, the confidence interval will include (or cover) the true mean (i.e., parameter). Confidence intervals are too often erroneously interpreted as saying that there is a 95 % probability of the true mean being within the limit bounds.

Two Worked Examples

The benefits of working through a few common examples seem enormous. In what follows I offer two different ‘power analyses’ for common study designs: the first is a t-test for a difference between two group means, the second example considers an odds-ratio from a case-control study. I rely on the PASS software package for each analysis [12]. There are other programs that yield similar results and I do not mean to suggest PASS is the best. But I do rely on it personally and find it user-friendly.

Two points must be emphasized before proceeding: (1) power analyses are always tailored to a particular study design and null hypothesis and (2) use of existing software is beneficial, but if study risks are high then expert guidance is necessary.

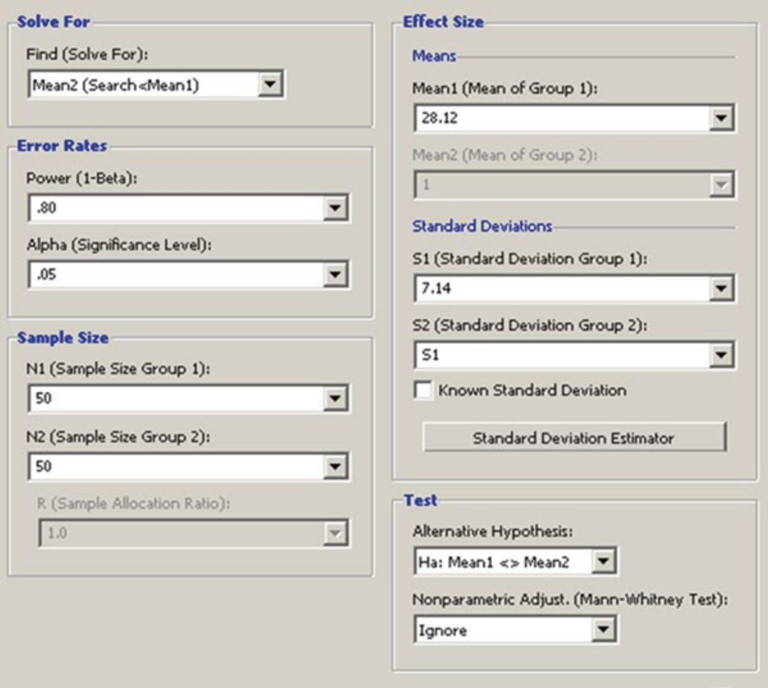
(Example 1) t-Test with Experimental Data

Imagine a simple randomized experiment where 50 subjects are given some treatment (the treatment group) and 50 subjects are not (the control or comparison group). Researchers might be interested in the difference in the mean outcome of some variable between groups. Perhaps we are interested in the difference in body mass index (BMI) between some diet regime and some control condition. Presume that it is known from pilot work and the existing literature that the mean BMI for the study population is 28.12 with a standard deviation of 7.14.

Since subjects were randomized to groups there is no great concern with confounding. A simple t-test between means will suffice for the analysis. Our null hypothesis is that the difference between means is nil; our alternative hypothesis is that the treatment group mean will be different (presumably but not necessarily less) than the control group mean.

Since we could only afford a total of N = 100 subjects, there is no reason to consider altering this. Additionally, we presume that in order to publish the results in a leading research journal we need 5 % Type I error and 20 % Type II error (or what is the same, 80 % Power). The question is, given the design and other constraints, how small an effect of the treatment can we detect? Inputting the necessary information into a software program is easy. The PASS screen for this analysis is shown in Fig. 15.3.

Notice that we are solving for ‘Mean 2 (Search < Mean 1)’ which implies that we are looking for the difference between our two sample means, where the second mean is less than the first or visa versa. Again, the alternative hypothesis is that our treatment group BMI mean will be different from the control groups, which is a non-directional or two-sided test. The specification here merely adds a sign (+ or −) to the estimated treatment effect. The question at hand is how small an effect can we minimally detect?



Solve For

Find (Solve For): Mean2 (Search<Mean1)

Error Rates

Power (1-Beta): .80

Alpha (Significance Level): .05

Sample Size

N1 (Sample Size Group 1): 50

N2 (Sample Size Group 2): 50

R (Sample Allocation Ratio): 1.0

Effect Size

Means

Mean1 (Mean of Group 1): 28.12

Mean2 (Mean of Group 2): 1

Standard Deviations

S1 (Standard Deviation Group 1): 7.14

S2 (Standard Deviation Group 2): S1

□ Known Standard Deviation

Test

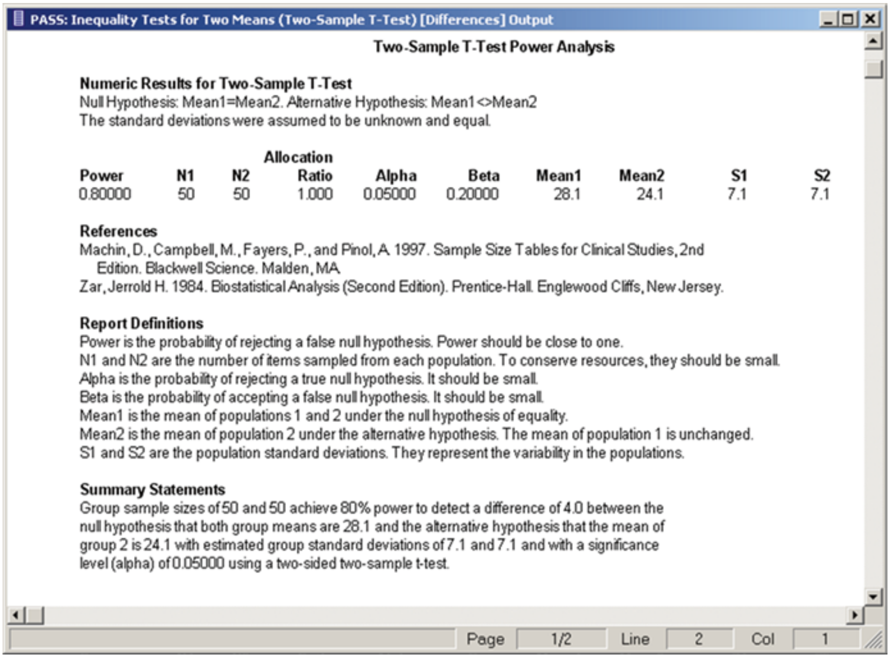
Alternative Hypothesis: Ha: Mean1 <> Mean2

Nonparametric Adjust. (Mann-Whitney Test): Ignore

Fig. 15.3 PASS input screen for t-test analysis

- We have given error rates for ‘Power’ to be 0.80 and our ‘Alpha (Significance)’ to be 0.05.
- The sample size we have is 50 for ‘N1 (sample size Group 1)’ and the same for ‘N2 (sample size Group 2)’. Again, we presume these are given due to budget constraints.
- The mean of group 1 ‘Mean1 (Mean of Group 1)’ is specific at 28.12, a value we estimated from our expertise and the existing literature. We are solving for the mean of group two ‘Mean2 (Mean of Group 2)’.
- The standard deviation of BMI also comes from the literature and is thought to be 7.14 for our target population (in the control or non-treatment arm). We assume that the standard deviation for the treatment arm will be identical to S1 or 7.14. Again, these are hypothetical values for this discussion only.
- The alternative hypothesis under investigation is that the means are unequal. This framework yields a 2-sided significance test, which is almost always indicated.

Clicking the ‘run’ button (top left) yields this PASS screen seen in Fig. 15.4, which is remarkably self-explanatory and detailed. The output shows that for 80 % Power, 5 % alpha or Type I error, two-sides significance test, 50 subjects per group, and a mean control-group BMI of 28.1 with a standard deviation of 7.1, we can expect to minimally detect a difference of 4.1 BMI units (28.1 − 44.1 = 4.0). To be clear, we have solved for Δ and it is 4.0. Given this design, we have an 80 % chance to detect


PASS: Inequality Tests for Two Means (Two-Sample T-Test) [Differences] Output
Page 1/2 Line 2 Col 1

```
                         Two-Sample T-Test Power Analysis

Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<>Mean2
The standard deviations were assumed to be unknown and equal.

                        Allocation
Power      N1     N2       Ratio     Alpha      Beta     Mean1     Mean2        S1        S2
0.80000    50     50       1.000   0.05000   0.20000      28.1      24.1       7.1       7.1

References
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd
   Edition. Blackwell Science. Malden, MA.
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions
Power is the probability of rejecting a false null hypothesis. Power should be close to one.
N1 and N2 are the number of items sampled from each population. To conserve resources, they should be small.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
Mean1 is the mean of populations 1 and 2 under the null hypothesis of equality.
Mean2 is the mean of population 2 under the alternative hypothesis. The mean of population 1 is unchanged.
S1 and S2 are the population standard deviations. They represent the variability in the populations.

Summary Statements
Group sample sizes of 50 and 50 achieve 80% power to detect a difference of 4.0 between the
null hypothesis that both group means are 28.1 and the alternative hypothesis that the mean of
group 2 is 24.1 with estimated group standard deviations of 7.1 and 7.1 and with a significance
level (alpha) of 0.05000 using a two-sided two-sample t-test.
```

Fig. 15.4 PASS output for t-test power analysis

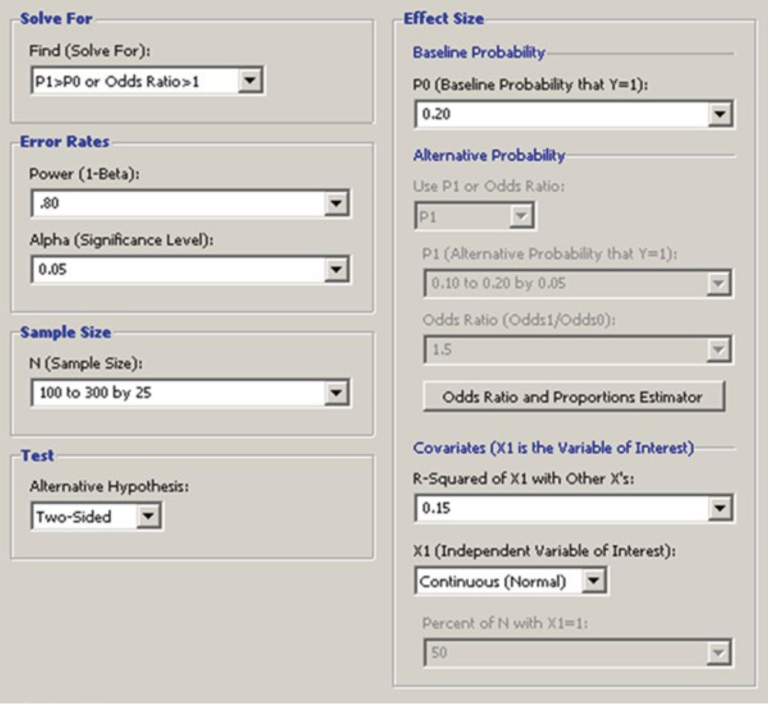
a 4.0 unit difference in BMI if in fact that difference exists. If our treatment actually has a larger impact on BMI, we will have more power to detect it.

If this change of 4.0 BMI units between treatment groups is thought to be possible and is clinically meaningful, then we have a well-designed study. If we can only hope for a 2.1 unit decrease in BMI from the intervention, then we are underpowered and should alter the study design. Possible changes include more subjects and or reducing the standard deviation of the outcome measure BMI, presumably by using a more precise instrument, or perhaps stratifying the analysis.

It is worth noting that more experienced users may examine the range of minimal detectable differences possible over a range of sample sizes or a range of possible standard deviations. Such 'sensitivity' analyses are very useful for both investigators and peer-reviewers.

(Example 2) Logistic Regression with Case-Control Data

The second example is for a (hypothetical) case-control study analyzed with a logistic regression model. Here again we navigate to the correct PASS input screen (Fig. 15.5) and input our desired parameters:



Solve For

Find (Solve For): P1>P0 or Odds Ratio>1

Error Rates

Power (1-Beta): .80

Alpha (Significance Level): 0.05

Sample Size

N (Sample Size): 100 to 300 by 25

Test

Alternative Hypothesis: Two-Sided

Effect Size

Baseline Probability

P0 (Baseline Probability that Y=1): 0.20

Alternative Probability

Use P1 or Odds Ratio: P1

P1 (Alternative Probability that Y=1): 0.10 to 0.20 by 0.05

Odds Ratio (Odds1/Odds0): 1.5

Covariates (X1 is the Variable of Interest)

R-Squared of X1 with Other X's: 0.15

X1 (Independent Variable of Interest): Continuous (Normal)

Percent of N with X1=1: 50

Fig. 15.5 PASS input screen

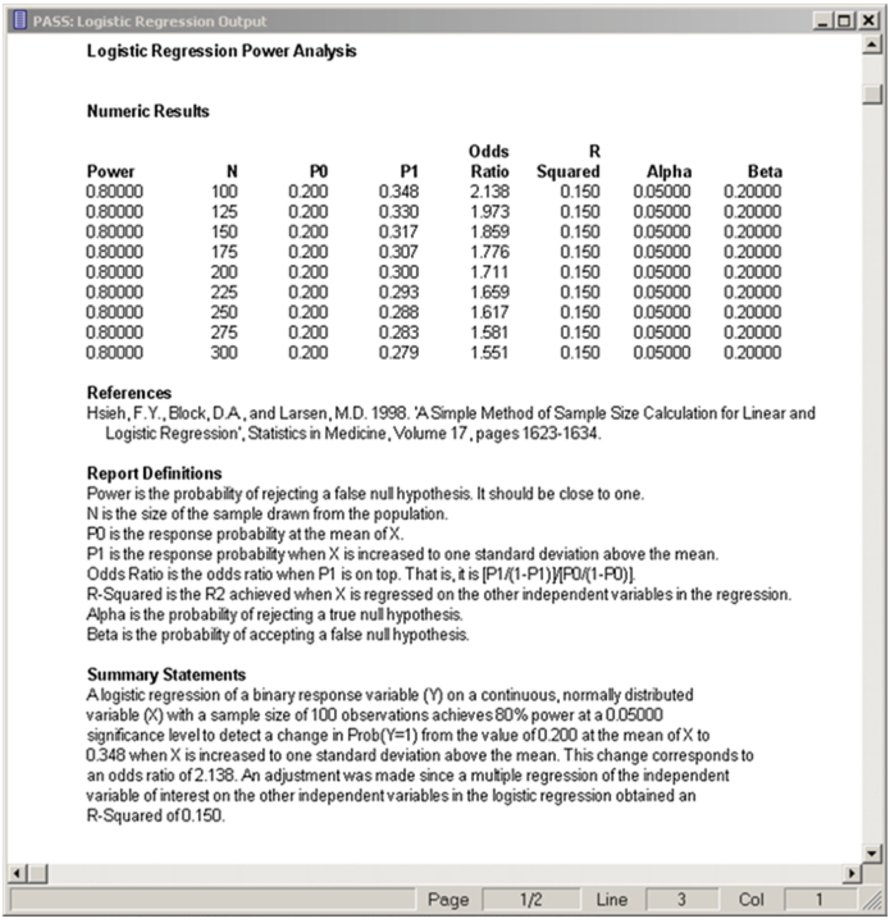
- Solve for an odds-ratios, expecting the exposure to have a positive impact on the outcome measure; in other words OR > 1.0
- Power = 80 % and Type I or alpha error = 5 %
- Let sample size vary from N = 100 to N = 300 by 25 person increments
- Two sided hypothesis test
- Baseline probability of exposure (recall this is case-control) of 20 %

And the explanatory power of confounders included in the model is 15 %.

But given the *range* of sample size values we specified, the output screen is shown in Fig. 15.6.

Given the null hypothesis of no effect (OR = 1.0), it is easy to see that the minimum detectable difference of exposure in this case-control study with N = 100 subject is 0.348 – 0.200 = 0.148, which is best understood as an OR = 2.138. With 300 subjects the same parameter falls to 1.551. As expected, increasing sample size (threefold) decreases the smallest effect one can expect to detect. Again, practically speaking, the smaller the better.

One can copy the actual values presented into a spreadsheet program (e.g., Microsoft Excel) and graph the difference in odds-ratios (that is, Δ) as a function of sample size. Reviewers tend to prefer such 'sensitivity' analyses. When it comes to such simple designs, this is about all there is to it, save for proper interpretation of course.



Logistic Regression Power Analysis

Numeric Results

Power	N	P0	P1	Odds Ratio	R Squared	Alpha	Beta
0.80000	100	0.200	0.348	2.138	0.150	0.05000	0.20000
0.80000	125	0.200	0.330	1.973	0.150	0.05000	0.20000
0.80000	150	0.200	0.317	1.859	0.150	0.05000	0.20000
0.80000	175	0.200	0.307	1.776	0.150	0.05000	0.20000
0.80000	200	0.200	0.300	1.711	0.150	0.05000	0.20000
0.80000	225	0.200	0.293	1.659	0.150	0.05000	0.20000
0.80000	250	0.200	0.288	1.617	0.150	0.05000	0.20000
0.80000	275	0.200	0.283	1.581	0.150	0.05000	0.20000
0.80000	300	0.200	0.279	1.551	0.150	0.05000	0.20000

References
Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. 'A Simple Method of Sample Size Calculation for Linear and Logistic Regression', Statistics in Medicine, Volume 17, pages 1623-1634.

Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population.
P0 is the response probability at the mean of X.
P1 is the response probability when X is increased to one standard deviation above the mean.
Odds Ratio is the odds ratio when P1 is on top. That is, it is [P1/(1-P1)]/[P0/(1-P0)].
R-Squared is the R2 achieved when X is regressed on the other independent variables in the regression.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis.

Summary Statements
A logistic regression of a binary response variable (Y) on a continuous, normally distributed variable (X) with a sample size of 100 observations achieves 80% power at a 0.05000 significance level to detect a change in Prob(Y=1) from the value of 0.200 at the mean of X to 0.348 when X is increased to one standard deviation above the mean. This change corresponds to an odds ratio of 2.138. An adjustment was made since a multiple regression of the independent variable of interest on the other independent variables in the logistic regression obtained an R-Squared of 0.150.

Fig. 15.6 PASS output screen

Conclusions

Sample size and statistical power are important issues for clinical research and it seems clinical researchers continue to struggle with the basic ideas. Accordingly, this chapter has aimed to introduce some fundamental concepts too often ignored in the more technical (i.e., precise) literature. Abundant citations are offered for those seeking more information or insight.

In closing, five points merit emphasis. First, sound inference comes from well-designed and executed studies. Planning is the key. Second, power analyses are always directly linked to a particular design and analysis (i.e., null hypothesis). General power calculations are simply not helpful, correct, and may even lead to disaster. Third, while used throughout this discussion, I emphasize that I do not advocate for the conventional 80 % power and 5 % Type I error. I simply use these above as common examples. Error rates should be carefully considered. Power analyses are properly done in the planning stage of a study. Retrospective power analyses are to be avoided [13]. Fourth, assumptions of planned analyses are key. Multiple comparisons and multiple hypothesis tests undermine power calculations and assumptions. Further, interactive model specification (i.e., data mining) invalidates assumptions. Finally, cautions of when to consult a statistical expert are important, especially when research places subjects at risk.

For greater technical precision and in-depth discussion, interested readers are encouraged to examine the following texts, ordered from simplest to more demanding discussions: [1, 14, 15] A solid and more technical recent but general discussion is by Maxwell, Kelley and Rausch [16]. Papers more tailored to particular designs include Oakes and Feldman [17], Feldman and McKinlay [18], Armstrong; [19] Greenland; [20] Self and Mauritsen [21]. Of note is that Bayesian approaches to inference continue to evolve and merit careful study if not adoption by practicing statisticians [3]. Because the approach incorporates *a priori* beliefs and is focused on decision-making under uncertainty, the Bayesian approach to inference is actually a more natural approach to inference in epidemiology and clinical medicine.

Acknowledgements This chapter would not have been possible without early training by Henry Feldman and the outstanding comments and corrections of Peter Hannan.

References

1. Baussell RB, Li Y-F. Power analysis for experimental research: a practical guide for the biological, medical and social sciences. New York: Cambridge University Press; 2002. p. ix.
2. Herman A, Notzer N, Libman Z, Braunstein R, Steinberg DM. Statistical education for medical students – concepts are what remain when the details are forgotten. Stat Med. 2007;26:4344–51.
3. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. Clin Trials. 2005;2:295–300; discussion 301–4, 364–78.
4. Berry DA. Bayesian statistics. Med Decis Mak. 2006;26:429–30.
5. Browne RH. Using the sample range as a basis for calculating sample size in power calculations. Am Stat. 2001;55:293–8.
6. Bloom HS. Minimum detectable effects: a simple way to report the statistical power of experimental designs. Eval Rev. 1995;10:547–56.
7. Greenland S. Power, sample size and smallest detectable effect determination for multivariate studies. Stat Med. 1985;4:117–27.
8. Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology. 2001;12:291–4.
9. Savitz DA, Tolo KA, Poole C. Statistical significance testing in the American Journal of Epidemiology, 1970–1990. Am J Epidemiol. 1994;139:1047–52.
10. Sterne JA. Teaching hypothesis tests – time for significant change? Stat Med. 2002;21:985–94; discussion 995–9, 1001.
11. Greenland S. On sample-size and power calculations for studies using confidence intervals. Am J Epidemiol. 1988;128:231–7.
12. Hintz J. PASS 2008, statistical analysis and sample size software. NCSS, LLC. www.ncss.com. Accessed 22 Oct 2013.

13. Hoenig JM, Heisey D. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55:19–24.
14. Chow S-C, Shao J, Wang H. Sample size calculations in clinical research. New York: Marcel Dekker; 2003.
15. Lipsey M. Design sensitivity: statistical power for experimental research. Newbury Park: Sage; 1990.
16. Maxwell SE, Kelly K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. Annu Rev Psychol. 2008;59:537–63. doi:10.1146/annurev.psych.59.103006.093735.
17. Oakes JM, Feldman HA. Statistical power for nonequivalent pretest-posttest designs. The impact of change-score versus ANCOVA models. Eval Rev. 2001;25:3–28.
18. Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. Stat Med. 1994;13:61–78.
19. Armstrong B. A simple estimator of minimum detectable relative risk, sample size, or power in cohort studies. Am J Epidemiol. 1987;126:356–8.
20. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med. 1983;2:243–51.
21. Self SG, Mauritsen RH. Power/sample size calculations for generalized linear models. Biometrics. 1988;44:79–86.

Chapter 16
Association, Cause, and Correlation

Stephen P. Glasser and Gary Cutter

The star of the play is the effect size i.e. what you found
The co-star is the effect size's confidence interval i.e. the precision that you found
If needed, supporting cast is the adjusted analyses i.e. the exploration of alternative explanations
With a cameo appearance of the p value, which, although its career is fading, insisted upon being included
Do not let the p value or an F statistic or a correlation coefficient steal the show, the effect size must take center stage!
But remember it takes an entire cast to put on a play!

Abstract Anything one measures can become data, but only those data that have meaning can become information. Information is almost always useful; data may or may not be. This chapter will address the various ways one can measure the degree of association between an exposure and an outcome and will include a discussion of relative and absolute risk, odds ratios, number needed to treat, and related measures. In addition, this chapter will introduce the concept of causal inference.

Keywords Probability • Chi-square test • Relative risk • Attributable risk • Relative risk reduction • Number needed to treat • Correlation • Regression • Causation

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

G. Cutter, Ph.D.
School of Public Health, University of Alabama at Birmingham,
Birmingham, AL, USA

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_16,

Introduction

Anything one measure's can become data, but only those data that have meaning can become information. Information is almost always useful; data may or may not be. Types of data include dichotomous, categorical, and continuous. For finding associations in clinical research data, there are several tools available. For categorical analyses one can compare relative frequencies or proportions, cross classifications (grouped according to more than one attribute at the same time) offering three different kinds of percentages (row, column, and joint probabilities), and to assess whether these are different from what one might expect by chance: chi square tests. When comparing continuous measures one can use correlation, regression, analysis of variance (ANOVA), and survival analyses. The techniques for continuous variables can also accommodate categorical data into their assessments.

Relative Frequencies and Probability

Let's address relative frequencies first, or how often something appears relative to all results. The simplest relative frequency can be a probability, a rate (a numerator divided by what is in the numerator plus what is not in the numerator) i.e. A/A + B (Influenza fatality rate: those who are infected with influenza and die denoted by A divided by those infected who die (A) plus those infected who recover (B)). In contrast, a ratio is written as A/B, where the numerator is not part of the denominator. Examples of rates are the probability of a 6 on the throw of a die (there is one 6 and 5 other 'points' on the die), or the probability of a winning number in roulette. Three key concepts in probability and associations are: joint probability, marginal probability, and conditional probability (i.e. probability of A occurs given B has already occurred). Figure 16.1 diagrams these three types of probabilities. These concepts are key to cross classifications of variables.

Dependence is another way of saying association, and two events are dependent if the probability of A and B (A & B) occurring is not equal to the probability of A times the probability of B. If the probability of A & B is equal to the product of the probability of A times the probability of B the two events are said to be independent. For example, there are 4 suits in a deck of cards, thus, the probability of drawing a card that is a heart is ¼. There are 4 queens in a deck of cards, thus the probability of drawing a queen is 4/52. The probability of drawing the queen of hearts is ¼ times 4/52 = 1/52. Thus we can say that the suit of the card is independent of the face on the card. How does this apply to epidemiology and medical research? To illustrate, consider the 2 × 2 table shown in Table 16.1.

By applying the above to an exploration of the association of hormone replacement therapy (HRT) and deep venous thrombosis (DVT) from the following theoretic data we can calculate the joint, marginal, and conditional probability as seen in Table 16.2.

Fig. 16.1 Venn diagram of conditional vs. joint probability

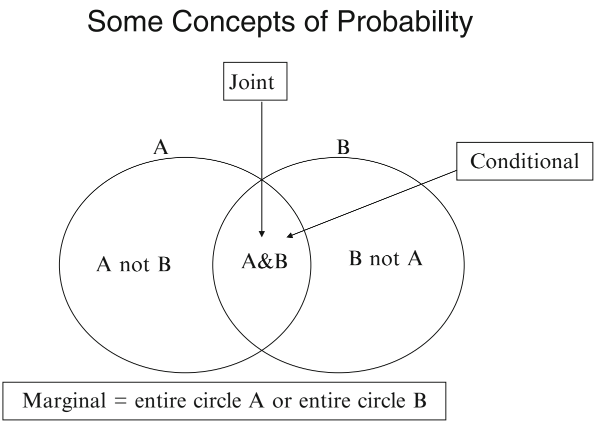
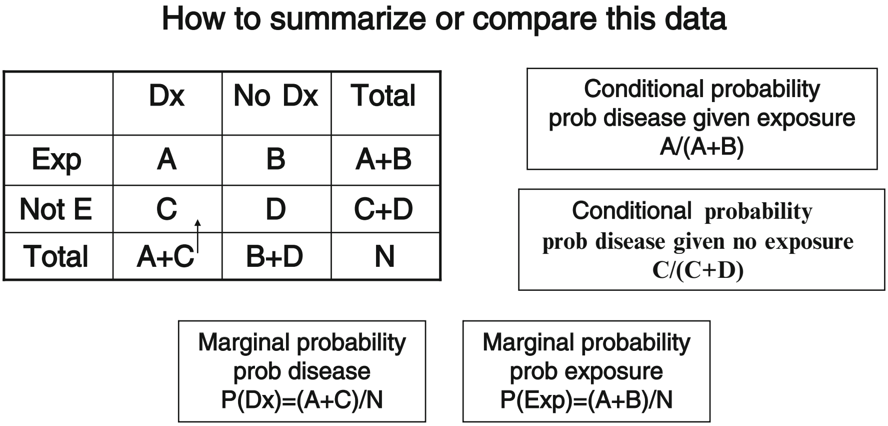

Some Concepts of Probability
Joint
A
B
Conditional
A not B
A&B
B not A
Marginal = entire circle A or entire circle B

Table 16.1 An example of summarizing data and defining marginal and conditional probability



How to summarize or compare this data

	Dx	No Dx	Total
Exp	A	B	A+B
Not E	C	D	C+D
Total	A+C	B+D	N

Conditional probability
prob disease given exposure
A/(A+B)

Conditional probability
prob disease given no exposure
C/(C+D)

Marginal probability
prob disease
P(Dx)=(A+C)/N

Marginal probability
prob exposure
P(Exp)=(A+B)/N

Table 16.2 An example of calculating outcome (Deep Vein Thrombosis-DVT) and exposure to Hormone Replacement Therapy (HRT)

The two by two table HRT and DVT

	DVT	No DVT	Total
On HRT	33	1,667	1,700
No HRT	27	2,273	2,300
Total	60	3,940	4,000

Marginal Probability of DVT $= 60/4{,}000$ On HRT $= 1{,}700/4{,}000$

$= 0.0150$ or 1.50 % $= 0.4250$ or 42.50 %

Conditional Probability of DVT given you are on HRT

$= 33/1{,}700 = 0.0194$ or 1.94 %

Conditional Probability of DVT given you're not on HRT

$= 27/2{,}300 = 0.0117$ or 1.17 %

Joint Prob of HRT and DVT

$= 33/4{,}000 = 0.0083$ or 0.83 %

To test the hypothesis of independence, we can use the above probability rules to determine how well the observed compares to the expected occurrence under the assumption that the HRT therapy is independent of the DVT. One way to do this is to use the chi square statistic (basically the observed value in any of the cells of our cross-classification minus the expected value squared, divided by the expected for each cell of our cross-classification table; and, add up these squared deviations to achieve a test statistical value). If the value that is calculated occurs by chance (found by comparing to the appropriate chi-square distribution table) is less than say 5 % we will reject the hypothesis that the row and column variables are independent, thereby implying that they are not independent i.e. an association exists. Any appropriately performed test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting a hypothesis. Typically, the hypothesis tested with chi square is whether or not two different samples (of people, texts, whatever) are different enough in some characteristic or aspect of their behavior that we can say from our sample that the populations from which our samples are drawn appear to be different from the expected behavior.

A non-parametric test (specific distributions of values are not specified a priori, some assumptions are made such as independent and identically distributed values) makes fewer assumptions about the form of the data occurring, but compared to parametric tests (like t-tests and analysis of variance, for example) it is less powerful or less likely to identify an association and, therefore, has less status in the list of statistical tests. Nonetheless, its limitations are also its strengths; thus, because chi- square is more ‘forgiving’ in the data it will accept, it can be used in a wide variety of research contexts.

Generalizing from Samples to Populations

Converting raw observed values or frequencies into percentages does allow us to more easily see patterns in the data, but that is all we can see, unless we make some additional assumptions about the data. Knowing with great certainty how often a particular drink is preferred in a particular group of 100 students is of limited use; we usually want to measure a sample in order to infer something about the larger populations from which our samples were drawn. On the basis of raw observed frequencies (or percentages) of a sample’s behavior or characteristics, we can make claims about the sample itself, but we cannot generalize to make claims about the population from which we drew our sample. To make assumptions about the population from which we drew our sample, we make some assumptions on how that sample was obtained and submit our results to quantification, so called inferential statistics; and, often to make inferences, a test of statistical significance. A test of statistical significance tells us how confidently we can generalize to a larger (unmeasured) population from a (measured) sample of that population (see the Chap. 18).

How does the chi square distribution and test statistic allow us to draw inferences about the population from observations on a sample? The chi-square statistic is what statisticians call an enumeration statistic. Rather than measuring the value of each of a set of items, a calculated value of chi-square compares the frequencies of various kinds (or categories) of items in a random sample, to the frequencies that are expected if the population frequencies are as hypothesized by the investigator. Chi square is often called a 'goodness of fit' statistic. That is, it compares the observed values to how well they fit what is expected in a random sample and what is expected under a given statistical hypothesis. For example, chi-square can be used to determine if there is a reason to reject the statistical hypothesis (i.e. the change that it arose from the underlying model given the expected frequencies is so low that we choose to assume the underlying model is incorrect). For example, we might want to know that the frequency in a random sample is consistent with items that come from a normal distribution. We can divide up the normal distribution into areas, calculate how many items would fall within those areas and compare to how many fall in those areas from the observed values.

Basically then, the chi square test of statistical significance is a series of mathematical formulas that compare the actual observed frequencies of some phenomenon (in a sample) with the frequencies we would expect. In terms of determining associations, we are testing that the fit of the observed data to that expected if there were no relationships at all between the two variables in the larger (sampled) population. The chi-square tests our actual results against the null hypothesis that the items were the result of an independent process and assesses whether the actual results are different enough from what might occur just by sampling error.

Chi Square Requirements

As mentioned before, chi square is a nonparametric test, that is it does not require the sample data to be more or less normally distributed (like parametric tests such as the t-tests do); although it relies on the assumption that the variable is sampled randomly from an appropriate population.

But, chi square, while forgiving, does have some requirements as noted below:

1. It must be assumed that the sample is randomly drawn from the population.
   As with any test of statistical significance, your data is assumed to be from a random sample of the population to which you wish to generalize your claims. While nearly never technically true, we make this assumption and must consider the implications of violating this assumption (i.e. a biased sample).
2. Data must be reported in raw frequencies (not percentages); one should only use the chi square when your data are in the form of raw frequency counts of things in two or more mutually exclusive and exhaustive categories. As discussed above, converting raw frequencies into percentages standardizes cell frequencies as if there were 100 subjects/observations in each category of the independent variable for comparability; but, this is not to be used in calculations of the chi square statistic. Part of the chi square mathematical procedure accomplishes this standardizing, so computing the chi square on percentages would amount to

standardizing an already standardized measurement and would always assume that there were 100 observations irrespective of the true number, thus in general, would give the wrong answer except when there are exactly 100 observations

3. Measured variables must be measured independently between people;
   That is, if we are measuring disease prevalence using sisters in the group, the measurement may not be an independent assessment, since there may be strong familial risk of the disease.
4. Values/categories on independent and dependent variables must be mutually exclusive and exhaustive (each person or observation can only go into one place)
5. Expected frequencies cannot be too small. The computation of the chi-square test involves dividing the difference between the observed and expected value squared by the expected value. If the expected value were to small, this calculation could wildly distort the statistic. A general rule of thumb is that the expected must be greater than 1 and not more than 20 % of the expected values and should be less than 5.

We will discuss expected frequencies in greater detail later, but for now remember that expected frequencies are derived from observed frequencies under an independence model.

Relative Risk and Attributable Risk (Fig. 16.2)

One of the more common measures of association is relative risk (RR). Relative Risk is the incidence of disease in one group compared to the other. As such it is used as a measure of association in cohort studies and RCTs. Said in other ways, RR is the risk of an event (or of developing a disease) in one group relative to another; or, it is a ratio of the probability of the event occurring in the exposed group versus the probability of an event occurring in the control (non-exposed) group.

$$RR = \frac{p_{\text{exposed}}}{p_{\text{control}}}$$

For example, if the probability of developing lung cancer among smokers was 20 % and among non-smokers 1 %, then the relative risk of cancer associated with smoking would be 20. Smokers would be twenty times as likely as non-smokers to develop lung cancer. Relative risk is used frequently in the statistical analysis of binary outcomes where the outcome of interest has relatively low probability. It is thus often an important outcome of clinical trials, where it is used to compare the risk of developing a disease say in people not receiving a new medical treatment (or receiving a placebo) versus people who are receiving a new treatment. Alternatively, it is used to compare the risk of developing a side effect in people receiving a drug as compared to the people who are not receiving the treatment (or receiving a placebo). A relative risk of 1 means there is no difference in risk between the two groups (since the null hypothesis is operative a RR implies no association between exposure and

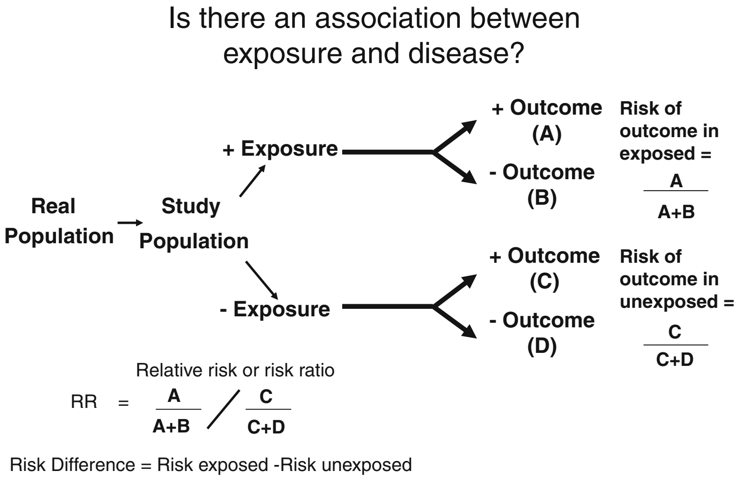

Is there an association between exposure and disease?
Real Population
Study Population
+ Exposure
- Exposure
+ Outcome (A)
- Outcome (B)
+ Outcome (C)
- Outcome (D)
Risk of outcome in exposed = $\frac{A}{A+B}$
Risk of outcome in unexposed = $\frac{C}{C+D}$
Relative risk or risk ratio
RR = $\frac{A}{A+B} / \frac{C}{C+D}$
Risk Difference = Risk exposed -Risk unexposed

Fig. 16.2 A schematic diagram of determining the association between exposure and disease

outcome) and the study then seeks to disprove that there is no association (the alternative hypothesis).

- A RR of <1 means the event is less likely to occur in the experimental group than in the control group.
- A RR of >1 means the event is more likely to occur in the experimental group than in the control group.

In the standard or classical hypothesis testing framework, the null hypothesis is that RR = 1 (the putative risk factor has no effect). The null hypothesis can be rejected in favor of the alternative hypothesis that the factor in question does affect risk (if the confidence interval for RR excludes 1, a so-called two sided test, since the RR can be less than one or greater than 1). A RR of >2 suggests that the intervention is 'more likely than not' (also a legal term) responsible for the outcome. Since RR is a measure of incident cases, RR cannot be used in case control studies because case control studies begin with the identification of existent cases, and matches controls only to cases. With the RR one needs to know the incidence in the unexposed group, and because the number of nonexposed cases is under the control of the investigator, there isn't an accurate denominator from which to compute incidence. This latter issue also prevents the use of RR even in prospective case-control studies (see discussion of case control study designs). In such situations, we use an approximation to the RR, called the Odds Ratio (OR, discussed below), which when incident rates are under 10 % the comparison of RR and OR works very well.

Whether a given RR can be considered statistically significant is dependent on whether the relative difference between conditions are being compared, and the amount of measurement and "noise" associated with the measurement. In other words, the amount of confidence that one has that a given RR is non-random is dependent on the effect size, the amount of noise, and the sample size of the study. A small effect size may be important in some situations (i.e. may be clinically important) and whether a given treatment is worthy is further influenced by its risks, benefits and cost.

Attributable risk (AR) is a measure of the excess risk that can be attributed to an intervention, above and beyond that which is due to other causes. When the AR exceeds 50 %, it is about equivalent to a RR >2. AR = incidence in the exposed group minus incidence in the unexposed divided by the incidence in the exposed. Thus, if the incidence of disease in the exposed group is 40 % and in the unexposed is 10 %, the proportion of disease that is attributable to the exposure is 75 % (30/40). That is, 75 % of the cases are due to the exposure. By the way, 'attributable' does not mean causal.

Odds Ratio

Another common measure of association is the odds ratio (OR). As noted above, it is used in case control studies as an alternative to the RR. The OR is a way of comparing whether the probability of a certain event is the same for two groups, with an OR of 1 implying that the event is equally likely in both groups (as is true with the RR). The odds of an event occurring, is a ratio; the occurrence of the event divided by the lack of its occurrence (Table 16.3). Commonly one hears in horse racing that the horse has 4 to 1 odds of winning. This means that if the race were run four times, this horse is expected to win three times and lose one time. Another horse may have 2 to 1 odds. The odds ratio between the two horses would be 3/1 divided by 2/1 or 1.5. Thus, the odds ratio of the first horse winning to the second is 1.5.

$$\text{Odds ratio} = \left(\text{Pi}/\left(1-\text{Pi}\right)\right)/\left(\text{Pc}/\left(1-\text{Pc}\right)\right)$$

The odds ratio approximates the relative risk only when the probability of endpoints (event rate or incidence) is lower than 10 %. Above this threshold, the odds ratio will overestimate the relative risk. It is easy to verify the 'lower than 10 %' rule. The relative risk from the odds ratio is:

$$\text{Relative risk} = \text{Odds ratio}/\left(1+\text{Pc}*\left(\text{Odds ratio}-1\right)\right)$$

Table 16.3 An example of how to calculate the odds ratio

The odds ratio		
	Dx	**No Dx**
Exp	A	B
Not E	C	D

The odds of cancer given exposure is A:B or A/B

The odds of cancer given no exposure is C:D or C/D

The odds ratio of cancer is: A/B divided by C/D

O.R. = AD/BC

Thus, for ORs larger than 1, the RR is less than or equal to the OR. The odds ratio has much wider use in statistics. Because the log of the odds ratio is estimated as a linear function of the explanatory variables, statistical models of the odds ratio often reflect the underlying mechanisms more effectively. When the outcome under study is relatively rare, the OR and RR are very similar in terms of their measures of association, but as the incidence of the outcome under study increases, the OR will underestimate the RR [1].

Since relative risk is a more intuitive measure of effectiveness, the distinction above is important, especially in cases of medium to high event rates or probabilities. If action A carries a risk of 99.9 % and action B a risk of 99.0 % then the relative risk is just slightly over 1, while the odds associated with action A are almost ten times higher than the odds with B. In medical research, the odds ratio is used frequently for case-control studies and retrospective studies because it can be obtained more easily and with less cost than studies which must estimate incidence rates in various risk groups. Relative risk is used in randomized controlled trials and cohort studies, but requires longitudinal follow-up and thus is more costly and difficult to obtain [2].

Relative Risk Reduction (RRR) and Absolute Risk Reduction (ARR) and Number Needed to Treat (NNT)

The RRR is simply 1 –RR times 100, and is the difference in event rates between two groups (e.g. a treatment and control group). Let's say you have done a trial with 100 patients in the intervention group and 100 patients in the control group, and there are 30 events in the former and 40 in the latter. The RRR is 25 % (i.e. 30/100 compared to 40/100) or a 10 % absolute reduction. The absolute risk reduction ARR is just the difference in the incidence rates. So the ARR above is .40 minus 0.30 or .10, a difference of 10 cases. But, what if in another trial we see events of 20 % in the control group vs.15 % in the intervention group. The RRR is 5/20 or 25 % while the ARR is only 5 %.

Absolute risk reduction (ARR) is another possible measure of association that is becoming more common in reporting clinical trial results of a drug intervention. Its inverse is called the number needed to treat or NNT. The ARR is computed by subtracting the proportion of events in the control group from the proportion of events in the intervention group. NNT is 1/ARR and is a relative measure of how many patients need to be treated to prevent 1 outcome event (in a specified time period). If there are 5 out of 100 outcomes in the intervention group (say you are measuring strokes with BP lowering in the experimental group over12 months of follow-up) and 30/100 in the control group, the ARR is $.30 - .05 = .25$, and the NNT is 4 (1/.25), that is for every four patients treated for a year (in the period of time of the study usually amortized per year) 1 stroke would be prevented (this, by the way, would be a highly effective intervention). Table 16.4 summarizes the various measures of association and, Table 16.5 summarizes the calculations of therapeutic effect from the example above.

Table 16.4 Formulas for common measures of association

Measure of effect	Formula
Relative Risk (RR)	Event rate in intervention group/event rate in control group
Relative Risk Reduction (RRR)	1- RR or absolute risk reduction/event rate in control group
Absolute Risk Reduction (ARR)	Event rate in intervention group – event rate in control group
Number Needed to Treat (NNT)	1+ARR

Table 16.5 Calculations from example in text and summarized below for Drug X

Parameter	Treatment drug X	Control treatment
Events/N=Rate	5/100=0.05	30/100=.30
Relative risk	.05/.30=0.17	.40/.05=6
Odds ratio	5×70/30×95=0.12	30×95/5×70=8.1
Absolute risk reduction	.30−.05=0.25	
Number Needed to Treat (NNT)	1/(.30−.05)=4	

Drug X (5 events) compared to control (30 events) in 100 patients per group

Table 16.6 A potentially misleading measure of association

	Annual mortality rate lung cancer	Annual mortality rate coronary heart disease
Smokers	140	669
Non-smokers	10	413
Relative risk	14	1.6
Attributable risk	130/100,000/year	256/100,000/year

Table 16.7 Example of marked differences of NNT (treatment with captopril to prevent one death)

	Control deaths	Intervention deaths	RR	NNT
SAVE trial	275/1,115 (24.7 %)	228/1,116 (20.4 %)	0.828	24
ISIS 4	2,231/29,022 (7.69 %)	2,088/29,028 (7.19 %)	0.936	201

The main issue in terms of choosing any statistic, but specifically a measure of association, is to not use a measure of association that could potentially mislead the reader. An example of how this can happen is shown in Table 16.6. In this example the RR of 14 for annual lung cancer mortality rates is compared to the RR of 1.6 for the annual mortality rate of CAD. However, at a population level, the mortality rate for CAD per 100,000 is almost twice that of lung cancer. Thus, while the RR is enormously higher, the impact of smoking on CAD in terms of disease burden (ARR) is nearly double. A further example from the literature is shown in Table 16.7, where the NNT to avoid one death with captopril is markedly different between two studies while the RRs were similar. One can also compute the NNH (number needed to harm), an important concept to carefully represent the downside of treating along with the upsides. The NNH is computed by subtracting the proportion of adverse events in the control and intervention group per Table 16.8.

Some experts argue that attention should also be paid to the absolute event rate observed in a trial and if it is not what is expected, questions about the trial and its

Table 16.8 Example of NNH

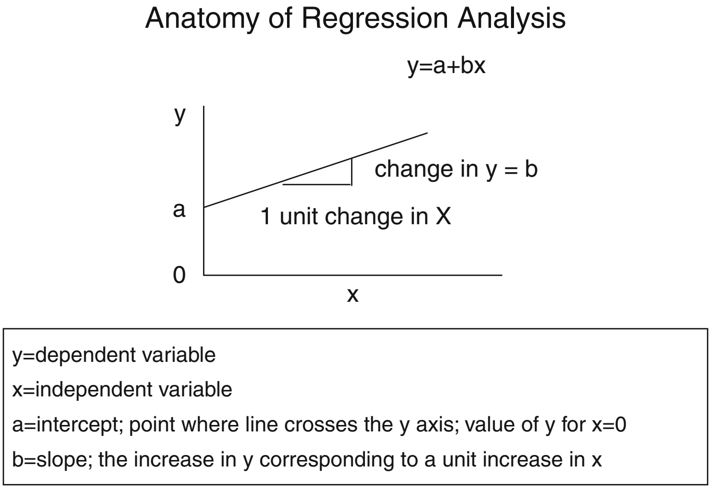
Adverse event	Finasteride (%)	Control (%)	NNH
Impotence	13.2	8.8	23
Decreased libido	9	6	33

Number needed to treat to result in one adverse event

results should be raised. Psaty and Prentice, compared the MI rates in control groups of six studies and pointed out the disparities between them also noting that the differences did not seem to be explained by differences in baseline risk factors [3]. They further suggested that the differences were more likely due to variations in event ascertainment or other study procedures. Indeed, they stated that "*the complete and accurate ascertainment of events in these trials seems key to the interpretation of their results and provides confidence about efforts to translate related new interventions into practice*." They point out that this is particularly problematic in non-inferiority trials and that "*without an explanation for the deviation from anticipated event rates, it may be unclear whether the findings are free from bias or whether their interventions merit widespread dissemination.*"

Correlations and Regression

Other methods of finding associations are based on the concepts above, but using methods that afford the incorporation of other variables and include such tools as correlations and regression (e.g. logistic, linear, non-linear, least squares regression line, multivariate or multivariable regression, etc.). We use the term regression to imply a co-relationship, and the term correlation to show relatedness of two or more variables. Linear regression investigates the linear association between two continuous variables. Linear regression gives the equation of the straight line that best describes an association in terms of two variables, and enables the prediction of one variable from the other. This can be expanded to handle multiple variables (multivariable regression). In general, regression analysis examines the dependence of a random variable, called the dependent or response variable, on other random or deterministic variables, called independent variables or predictors. The mathematical model of their relationship is known as the regression equation. This is an extensive area of statistics and in its fullest forms are beyond the scope of this chapter. Well known types of regression equations are linear regression for continuous responses, the logistic regression for discrete responses, and nonlinear regression. Besides dependent and independent variables, the regression equations usually contain one or more unknown regression parameters, which are to be estimated from the given data in order to maximize the quality of the model. Applications of regression include curve fitting, forecasting of time series, modeling of causal relationships and testing scientific hypotheses about relationships between variables. A graphical depiction of regression analysis is shown in Fig. 16.3. Correlation is the tendency for one variable to change as the other variable changes (it is measured by rho).



y=dependent variable
x=independent variable
a=intercept; point where line crosses the y axis; value of y for x=0
b=slope; the increase in y corresponding to a unit increase in x

Fig. 16.3 A diagram of regression analysis

Correlation, also called correlation coefficient, indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence, that is, knowledge of one variable better informs an investigator of the expected results of the dependent variable than not considering this covariate. Correlation does not imply causation, but merely that additional information is provided about the dependent variable when the covariate (independent variable) is known. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of data. The rate of change of one variable tied to the rate of change of another is known as a slope. The correlation coefficient and the slope of the regression line are functions of one another, and a significant correlation is the same as a significant regression. You may have heard of a concept called the r-squared. We talk of r-squared (r^2) as the percent of the variation in one variable explained by the other. This means that if we compute the variation in the dependent variable by taking each observation, subtracting the overall mean and summing the squared deviations and dividing by the sample size we arrive at our estimated variance. To assess the importance of the covariate we compute a 'regression' model using the covariate, and assess how well our model explains the outcome variable. We compute an expected value based on the regression model for each outcome. Then we assess how well our observed outcomes fit our expected. We compute the observed minus the expected, called the residual or unexplained portion and find the variance of these residuals. The ratio of the variance of residuals to the variation in the outcome variable overall is the proportion of unexplained variance and 1 minus this ratio is the R-squared or proportion of variance explained.

A number of different coefficients are used for different situations. The best known is the Pearson product-moment correlation coefficient, which is easily obtained by standard formulae. Geometrically, if one thinks of a regression line, it is a function of the angle that the regression line makes with a horizontal line parallel to the x-axis. Importantly, it should be realized that correlation can measure precision and/or reproducibility, but does not measure accuracy or validity.

Causal Inference

An association (or a correlation) does not imply causation. In an earlier chapter, various clinical research study designs were discussed, and the differing 'levels of scientific evidence' that are associated with each were reviewed. A comparison of study designs is complex, with the metric being that the study design providing the highest level of scientific evidence (usually experimental studies) is the one that yields the greatest likelihood of cause and effect between the exposure and the outcome. The basic tenet of science is that it is almost impossible to prove an association or cause, but it is easier to disprove it. Causal effect focuses on outcomes among exposed individuals, but what would have happened had they not been exposed? The outcome among exposed individuals is called the factual outcome. To draw inferences, exposed and non-exposed individuals are compared. Ideally, one would use the same population expose them, observe the result and then go back in time and repeat the same experiment among the same individuals but without the exposure to observe the counterfactual outcome. Randomized clinical trials attempt to approximate this ideal by using randomly assigned individuals to groups (to avoid any bias in assignment) and observe the outcomes. Because the true ideal experiment is impossible, replication of results with multiple studies is the norm. Another basic tenet is that even when the association is statistically significant, association does not denote causation. Causes are often distinguished into two types: Necessary and Sufficient.

Necessary Causes

If x is a necessary cause of y; then the presence of y necessarily implies the presence of x. The presence of x, however, does not imply that y will occur.

Sufficient Causes

If x is a sufficient cause of y, then the presence of x necessarily implies the presence of y. However, another cause z, may alternatively cause y. Thus the presence of y does not imply the presence of x.

The majority of these tenets and related ones (Koch's postulates, Bradford Hills tenets of causation) were developed with infectious diseases. There are more tenuous conclusions that emanate from chronic diseases.

Table 16.9 Five explanations for an association

Association	Basis	Type	Explanation
1. C → MI	Cause-effect	Real	Cause-effect
2. MI → C	Cart before horse	Real	Effect-cause
3. C ← x → MI	Confounding	Real	Effect-cause
4. C ≠ MI	Random error	Spurious	Chance
5. C ≠ MI	Systematic error	Spurious	Bias

Consider the finding of an association between coffee drinking and myocardial infarction (MI) (Table 16.9). Coffee drinking might be a 'cause' of the MI, as the finding of that association from a study might imply. However, some persons who have had an MI may begin to drink more coffee, in which case (instead of a cause-effect relationship) the association would be an 'effect-cause' relationship (also referred to as reverse causation).

The association between coffee drinking and MI might be mediated by some confounder (e.g., persons who drink more coffee may smoke more cigarettes, and it is the smoking that precipitates the MI) (Table 16.3). Finally, observed associations may be spurious as a result of chance (random error) or because of some systematic error (bias) in the study design. To repeat, in the first conceptual association in Table 16.9, coffee drinking leads to MI, so it could be casual. The second association represents a scenario in which MI leads to coffee drinking (effect-cause or reverse causation). An association exists, but coffee drinking is not causal of MI. In the third association, the variable x results in coffee drinking and MI, so it confounds the association between coffee drinking and MI. In the fourth and fifth associations, the results are spurious because of chance or some bias in the way in which the trial was conducted or the subjects were selected.

Thus, establishing cause and effect, is notoriously difficult and with chronic diseases has become even more of a challenge. In terms of an infectious disease – think about a specific flu – many flu-like symptoms occur without a specific viral agent, but for the specific flu, we need the viral agent to be present to produce the flu. What about Guillian-Barre Syndrome – it is caused by the Epstein Barr Virus (EBV), but the viral infection and symptoms have often occurred previously. It is only thru the antibodies to the EBV that this cause was identified. Further, consider the observation that smokers have a dramatically increased lung cancer rate. This does not establish that smoking must be a cause of that increased cancer rate: maybe there exists a certain genetic defect which both causes cancer and a yearning for nicotine; or even perhaps nicotine craving is a symptom of very early-stage lung cancer which is not otherwise detectable. In statistics, it is generally accepted that observational studies (like counting cancer cases among smokers and among non-smokers and then comparing the two) can give hints, but can never establish cause and effect. The gold standard for causation is the randomized experiment: take a large number of people, randomly divide them into two groups, force one group to smoke and prohibit the other group from smoking (obviously ethically unfeasible), then determine whether one group develops a significantly higher lung cancer rate.

Random assignment plays a crucial role in the inference to causation because, in the long run, it renders the two groups equivalent in terms of all other possible effects on the outcome (cancer) so that any changes in the outcome will reflect only the manipulation (smoking). Obviously, for ethical reasons the above experiment cannot be performed, but the method is widely applicable for other experiments. And our search for causation must try to inform us with data as similar to possible as the RCT.

Because causation cannot be proven, how does one approach the concept of 'proof'? The Bradford Hill criteria for judging causality remain the guiding principles as follows: the replication of studies in which the magnitude of effect is large; biologic plausibility for the cause-effect relationship is provided; temporality and a dose response exist; similar suspected causality is associated with similar exposure outcomes; and, systematic bias is avoided.

Deductive vs Inductive Reasoning

Drawing inferences about associations can be approached with deductive and inductive reasoning. An overly simplistic approach is to consider deductive reasoning as truths of logic and mathematics. Deductive reasoning is the kind of reasoning in which the conclusion is necessitated by, or reached from, previously known facts (the premises). If the premises are true, the conclusion must be true. This is distinguished from inductive reasoning, where the premises may predict a high probability of the conclusion, but do not ensure that the conclusion is true. That is, induction or inductive reasoning, sometimes called inductive logic, is the process of reasoning in which the premises of an argument are believed to support the conclusion but do not ensure it.

For example, beginning with the premises 'All ice is cold' and 'This is ice', you may conclude that 'This ice is cold'. An example where the premise being correct but the reasoning incorrect is 'this French person is rude so all French must be rude' (although some still argue that this is true). That is, deductive reasoning is dependent on its premises-a false premise can possibly lead to a false result, and inconclusive premises will also yield an inconclusive conclusion. We induce truths based on the interpretation of empirical evidence; but, we learn that these 'truths' are simply our best interpretation of the data at the moment and that we may need to change as new evidence is presented.

When using empirical observations to make inductive inferences, we have a greater ability to falsify a principle than to affirm it. This was pointed out by Karl Popper [3] in the late 1950s with his now classic example: if we observe swan after swan, and each is white, we may infer that all swans are white. We may observe 10,000 white swans and feel more confident about our inference. However, it takes but a single observation of a non-white swan to disprove the assertion. It is this Popperian view from which statistical inferences using the null hypothesis is born. That is we set our hypothesis that our theory is not correct, and then set out to disprove it. The p value is the probability (thus 'p'), that is the mathematical probability, that we would find a difference if the null hypothesis was true. Thus, the lower the probability of the finding, the more certain we can be in stating that we have falsified the null hypothesis.

Errors in making inferences about associations can also occur due to chance, bias, and confounding (See Chap. 17). Bias refers to anything that results in error i.e. compromises validity in a study. It is not (in a scientific sense) an intentional behavior, but rather it is an unintended consequence of a flaw in study design or conduct that affects an association. The two most common examples are selection bias (the inappropriate selection of study participants) and information bias (a flaw in measuring either the exposure group or disease group). These biases are the ‘achilles heel’ of observational studies which are essentially corrected for in randomized trials. However, randomized trials may restrict the study populations to a degree that also leads to selection biases. When an association exists, it must be determined whether the exposure caused the outcome, or the association is caused by some other factor (i.e. is confounded by another factor). A confounding factor is both a risk factor for the disease and a factor associated with the exposure. Some classify confounding as a form of bias. However, confounding is a reality that actually influences the association, although confounding can introduce bias (i.e. error) into the findings of a study. Confused with confounding is effect modification. Confounding and effect modification are very different in both the information each provides as well as what is done with that information. For confounding to exist, a factor must be unevenly distributed in the study groups, and as a result has influenced the observed association. Confounding is a nuisance effect, and the researchers main goal is to control for confounding and eliminate its effect (by stratification or multivariate analysis). In a statistical sense confounding is inextricably tied to the variable of interest, but in epidemiology we consider confounding a covariate. Effect modification is a characteristic that exists irrespective of study design or study patients. It is to be reported, not controlled (See further discussion of effect modification in Chap. 17).

Stratification is used to control for confounding, and to describe effect modification. If, for example, an association observed is stratified for age and the effect is not uniform across age groups, this suggests confounding by age. In contrast, if the observed association is not uniform, effect modification is present. For example, amongst premature infants, stratified by birth weight; 500–749 g, 750–999 g and 1,000–1,250 g, the incidence of intracranial hemorrhage (ICH) is vastly different across these strata, thus birth weight is an effect modifier of ICH.

Kohli and Cannon said “*because they (referring to each study design) are all different, the same language cannot be used to describe the results from distinct types of studies when drawing conclusions and characterizing the risk relationship between intervention and the outcome*”. The importance of matching language to type of evidence avoids the pitfalls of reporting outcome data [4]. One of the main points they were trying to make is that RCT’s come closest in approaching the issue of causality so it is appropriate to say that an intervention reduces clinical events, assuming that is what the data shows. The problem as they point out is when results

Table 16.10 The type of language suggested for different study designs

	Randomized trial	Observational study
Type of language		
Descriptive statement	"reduced risk by…"	"a lower risk was observed. There is a relationship, there is an association"
Descriptive noun	"relative risk reduction/benefit"	"difference in risk, risk ratio"
Verbs	"affected, caused, modulated risk, treatment resulted in…, reduced hazard"	"correlates with, is associated with"
Terms to avoid		"reduced risk, lowered risk, benefited"

From: Kohli and Cannon [4]

from an observational study are described as a treatment that provides direct clinical evidence of benefit or harm. In an accompanying editorial they point out that the descriptive statement for an RCT would be that the intervention reduced the risk by x amount but the same effect size in an observational study might best be stated as "a lower risk was observed, or there was a relationship or association with the exposure and outcome". Kohli and Cannon after pointing out that the language used in observational trials with hormone replacement therapy might have been overstated based upon the strength of the evidence conclude that "*therefore, it is important to be mindful of reporting these results with clear, accurate and consistent language that reflects the evidence being cited: for registry data, 'associated with' or 'relative risk ratio' are appropriate, whereas for the randomized data, 'risk reduction' is preferred*" (Table 16.10). Kohli and Cannon go on to point out that although the differences in the use of these terms may seem subtle, the implications can be significant, "in all types of observational studies, the authors should report the difference in outcome between two groups of patients descriptively; they cannot make conclusions about 'reductions' or 'increases' from this type of study." They point out that the use of "reduction or increase" implies causality in comparison to "correlated or associated with".

The p Value

Most often associations are thought of in terms of p-values. The significance level that is used most commonly is < 0.05 that represents the maximum probability that is tolerated for rejecting a hypothesis that is in fact true. For a further discussion see Chaps. 3 and 18. However, it should be realized that p values are influenced by sample size and variability in the measurement of outcomes and are favored by clinicians and clinical journals because they dichotomize an outcome into a "yes or no" answer. Epidemiologists prefer point estimates and confidence intervals because they feel this gives a much more representative picture.

Table 16.11 Statistics glossary. Some common statistical concepts and their use in analyzing experimental results

Term	Meaning	Common uses
Standard deviation (s.d.)	Typica1 difference between each value and the mean	Describes how broadly the sample va1ues are distributed
Standard error of the mean (s.e.m.)	Estimate of how variable the means will be if the experiment is repeated multiple times	Inferring where a population mean is likely to lie, or whether sets of samples are likely to come from the same population
Confidence interval (Cl; 95 %)	With 95 % confidence, the population mean will lie in this interval	To infer where the population mean lies, and to com pare two populations
Independent data	Values from separate experiments of the same type that are not linked	Testing hypotheses about the population
Replicate data	Values from experiments where everything is linked as much as possible	An internal check on the performance of an experiment
Sampling error	Variation caused by sampling part of a population	Can reveal bias in the data or problems with the conduct of the experiment

In summary, Vaux [5] noted that the number of papers that have basic statistical mistakes is alarming. To address this he provided a statistics glossary, as modified in Table 16.11.

References

1. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA. 1998;280:1690–1.
2. Relative Risk. In: Wikipedia. Available at: http://en.wikipedia.org/wiki/Relative_Risk
3. Psaty BM, Prentice RL. Variation in event rates in trials of patients with type 2 diabetes. JAMA. 2009;302:1698–700. doi:10.1001/jama.2009.1497.
4. Kohli P, Cannon CP. The importance of matching language to type of evidence: avoiding the pitfalls of reporting outcomes data. Clin Cardiol. 2012;35:714–7. doi:10.1002/clc.22066.
5. Vaux DL. Research methods: know when your numbers are significant. Nature. 2012;492: 180–1. doi:10.1038/492180a.

Chapter 17
Bias, Confounding, and Effect Modification (Interaction)

Stephen P. Glasser

You're like the Tower of Pisa-always leaning in one direction [1]

Abstract Bias, confounding, and random variation/chance are the reasons for a non-causal association between an exposure and outcome. This chapter will define and discuss these concepts so that they may be appropriately considered whenever one is interpreting the data from a study. Several types of common bias will be discussed (e.g. measurement bias, sampling bias, etc.) and effect modification (interaction) will be explained.

Keywords Bias • Confounding • Effect modification • Interaction

Introduction

Bias, confounding, and random variation/chance are alternate explanations for an observed association between an exposure and outcome. They represent a major threat to the internal validity of a study, and should always be considered when interpreting data. Whereas statistical bias is usually an unintended mistake made by the researcher; confounding is not a mistake; rather, it is an additional variable that can impact the outcome (negatively or positively; all or in part) separately from the exposure. Sometimes, confounding is considered to be a third major class of bias [2] (Table 17.1).

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_17,

Table 17.1 Alternative explanations (other that truth) for observed associations between exposure and outcome

Variable	Description	Correction
Bias	A systematic error in the design, recruitment, data collection or analysis	The most important design techniques for avoiding bias are blinding and randomization
Confounding	A situation in which the effect or association between an exposure and outcome is distorted by the presence of another variable	In the design phase (e.g. during randomization, restriction or matching) or in the analysis phase (by stratification, multivariable analysis and matching)
Effect modification	A variable that differentially (positively and negatively) modifies the observed effect of a risk factor on disease status	Statistical testing for interaction
Random chance	A chance effect (random variation)	Diminishes as sample size gets larger. A small p-value and a narrow CIs are reassuring signs against chance effect-the same cannot be said for bias and confounding

As will be further discussed, when a confounding factor is *known* or suspected, it can be controlled for (albeit never perfectly) in the design phase (e.g. during randomisation, restriction or matching) or in the analysis phase (by stratification, multivariable analysis and matching). The best that can be done about *unknown* confounders is to use a randomised design (see Chap. 3). Bias and confounding are not affected by sample size, but chance effect (random variation) diminishes as the sample size gets larger. A small p-value and a narrow confidence intervals (CIs) are reassuring signs against chance effect but the same cannot be said for bias and confounding [3].

Bias

Bias is a systematic error that results in an incorrect (invalid) estimate of a measure of association. That is, the term bias 'describes the systematic tendency of any factors associated with the design, conduct, analysis, and interpretation of the results of clinical research to make an estimate of a treatment effect deviate from its true value' [3]. Bias can either create or mask an association; that is, bias can give the appearance of an association when there really is none, or can mask an association when there really is one. Bias can occur with all study designs, be it experimental, cohort, or case-control; and, can occur in either the design phase of a study, or during the conduct of a study. For example, bias may occur from an error in the measurement of a variable; confounding involves an incorrect interpretation of an association even when there has been accurate measurement. Also, whereas adjustments can be made in the analysis phase of a study for confounding variables, bias cannot be controlled, at best; one can only suspect that it has occurred. The most important design techniques for avoiding bias are blinding and randomization.

Table 17.2 Examples of bias

We all know about the common problems in doing research	
Selecting study participants	Information biases
Selection bias	Recall bias
Non-respondent bias:	Reporting bias
Volunteer or referral bias	Family information bias
External validity	Measurement bias
Sampling bias	Misclassification bias
Ascertainment bias	Reporting bias
Prevalence-incidence bias	End-aversion bias
Berkson bias	Attention bias
Healthy worker effect	
Detection bias: The risk factor investigated itself may lead to increased	
Diagnostic	
Overmatching bias	

Berkson's bias is a type of selection bias which may occur in case-control studies which are based entirely on hospital studies

An example of systematic bias would be a thermometer that always reads three degrees colder than the actual temperature because of an incorrect initial calibration or labeling, whereas one that gave random values within five degrees either side of the actual temperature would be considered a random error [4]. If one discovers that the thermometer always reads three degrees below the correct value one can correct for the bias by simply making a systematic correction by adding three degrees to all readings. In other cases, while a systematic bias is suspected or even detected, no simple correction may be possible because it is impossible to quantify the error. The existence and causes of systematic bias may be difficult to detect without an independent source of information; the phenomenon of scattered readings resulting from random error calls more attention to itself from repeated estimates of the same quantity than the mutually consistent incorrect results of a biased system.

There are many types of bias (Table 17.2), but two common types are; selection and observation bias [5].

Selection Bias

Selection bias is the result of the approach used for subject selection. That is, when the sample in the study ends up being different from the target population, selection bias is a cause. Selection bias is more likely to be present in case-control or retrospective cohort study designs, because the exposure and the outcome have already occurred at time of subject selection. For a case-control study, selection bias occurs when controls or cases are more (or less) likely to be included in study if they have been exposed – that is, inclusion in the study is not independent of the exposure. The result of this is that the relationship between exposure and disease observed among study participants is different from relationship between exposure and

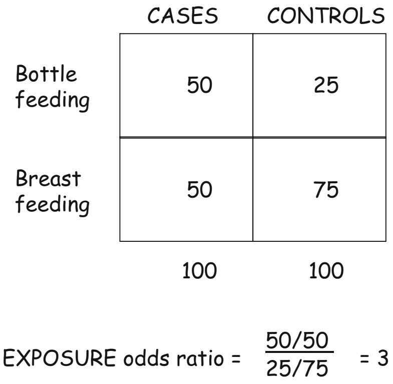


Fig. 17.1 Example of potential selection bias

	CASES	CONTROLS
Bottle feeding	50	25
Breast feeding	50	75
	100	100

$$\text{EXPOSURE odds ratio} = \frac{50/50}{25/75} = 3$$

disease in individuals who would have been eligible but were not included, thus the odds ratio from a study that suffers from selection bias will incorrectly represent the relationship between exposure and disease in the overall study population [2].

A biased sample is a statistical sample of a population in which some members of the population are less likely to be included than others. If the bias makes the estimation of population parameters impossible, the sample is a non-probability sample. An extreme form of biased sampling occurs when certain members of the population are totally excluded from the sample (that is, they have zero probability of being selected). For example, a survey of high school students to measure teenage use of illegal drugs will be a biased sample because it does not include home schooled students or dropouts. A sample is also biased if certain members are underrepresented or overrepresented relative to others in the population. For example, a “man on the street” interview which selects people who walk by a certain location is going to have an over-representation of healthy individuals who are more likely to be out of the home than individuals with a chronic illness. A biased sample causes problems because any statistic computed from that sample has the potential to be consistently erroneous [6]. Bias can lead to an over- or under-representation of the corresponding parameter in the population. Almost every sample in practice is biased because it is practically impossible to ensure a perfectly random sample. If the degree of under-representation is small, the sample can be treated as a reasonable approximation to a random sample. Also, if the group that is underrepresented does not differ markedly from the other groups in the quantity being measured, then a random sample can still be a reasonable approximation.

The word bias in common usage has a strong negative connotation, and implies a deliberate intent to mislead. In statistical usage, bias represents a mathematical property. While some individuals might deliberately use a biased sample to produce misleading results, more often, a biased sample is just a reflection of the difficulty in obtaining a truly representative sample [6].

Let’s take as an example the data shown in Fig. 17.1, which addresses the question of whether otitis media differs in bottle-feeding, as opposed to breast feeding. 100 infants with ear infection are identified among members of one HMO, and the controls are 100 infants in that same HMO without otitis. The potential bias is whether being included in the study as a control is not independent of the exposure, that is, they were not representative of the whole study population that produced the cases. In other words, one could ask the reason(s) that infants were being seen in an HMO in the first place and how many might have had undiagnosed otitis.

So, what are the solutions for selection bias? Little or nothing can be done to fix selection bias once it has occurred. Rather one needs to avoid it during the design and conduct of the study by, for example, using the same criteria for selecting cases and controls, obtaining all relevant subject records, obtaining high participation rates, and taking into account diagnostic and referral patterns of disease. But, almost always (perhaps always) one cannot totally remove selection bias from any study.

Observation Bias

While selection bias occurs as subjects enter the study, observation bias occurs after the subjects have entered the study. Observation bias is the result of incorrectly classifying the study participant's exposure or outcome status. There are several types of observation bias: recall bias, interviewer bias, loss to follow up, and differential and non-differential misclassification.

Recall bias occurs because participants with and without the outcome of interest do not report their exposure accurately (because they do not remember it accurately) and more importantly report the exposure differently (this can result in an over- or under-estimate of the measure of association). It is not that unlikely that subject's with an outcome might remember the exposure more accurately than subjects without an outcome, particularly if the outcome is a disease. Solutions for recall bias include using controls, who are themselves sick; and/or, using standardized questionnaires that obtain complete information and that mask subjects to the study hypothesis [7].

Whenever exposure information is sought, information is recorded and interpreted. If there is a systematic difference in the way the information is solicited, recorded, or interpreted, interviewer bias can occur. One solution to reduce interviewer bias is to mask interviewers, so that they are unaware of the study hypothesis and disease or exposure status of subjects, and to use standardized questionnaires or standardized methods of outcome (or exposure) ascertainment [8].

Loss to follow up is a concern in cohort and experimental studies if people who are lost to follow up differ from those that remain in the study (which is likely almost always the case). Bias results if subjects lost, differ from those that remain, with respect to both the outcome and exposure. The main solution for lost to follow up is to minimize its occurrence. Excessive numbers of subjects lost to follow up can seriously damage the validity of the study. (See discussion of lost to follow up in Chap. 3).

Misclassification bias occurs when a subject's exposure or disease status is erroneously classified. Two types of misclassification are non-differential (random) and differential (non random). Non-differential misclassification results in inaccuracies

Table 17.3 Questions to consider about bias

Could bias have occurred
Is bias actually present
Is bias large enough to distort the measure of association in an important way
Which direction is the distortion; toward or away from the null

with respect to disease classification that is independent of the exposure; or, with inaccuracies with respect to the exposure that are independent of disease. Non-differential misclassification makes the exposure and non- exposure groups more similar. The probability of misclassification may be the same in all study groups (non-differential misclassification) or may vary between groups (differential misclassification).

Measurement Bias

Let's assume that a true value does in fact exist. Both random and biological variation modifies that true value by the time the measurement is made. Performance of the instrument and observer bias, and recording and computation of the results further modifies the 'true value' and this now becomes the value used in the study. Reliability has to do with the ability of an instrument to measure consistently, repeatedly, and with precision and reproducibility. But, the fact is, that every instrument has some inherent imprecision and/or unreliability. This latter fact negatively impacts one of the main objectives of clinical research, to isolate between-subject variability from measurement variability. Measurement error is intrinsic to research.

In summary, in order to reduce bias, ask yourself these questions:' given the conditions of the study, could bias have occurred? Is bias actually present? Are consequences of the bias large enough to distort the measure of association in an important way? Which direction is the distortion, that is, is it towards the null or away from the null? (Table 17.3) [8].

Confounding

A confounding variable (confounding factor or confounder) is a variable that correlates (positively or negatively) with both the exposure and outcome. One, therefore, needs to control for these factors in order to avoid what is known as a type 1 error, which is a 'false positive' conclusion that the exposure is in a causal relationship with the outcome. Such a false relation between two observed variables is termed a spurious relationship. Thus, confounding is a major threat to the validity of inferences made about cause and effect, i.e. internal validity, as the observed effects should be attributed all or in part to the confounder rather than the outcome.

For example, assume that a child's weight and a country's gross domestic product (GDP) rise with time. A person carrying out an experiment could measure weight and GDP, and conclude that a higher GDP causes children to gain weight. However, the confounding variable, time, was not accounted for, and is the real cause of both rises [9]. By definition, a confounding variable is associated with both the probable cause and the outcome, and the confounder should not lie in the causal pathway between the cause and the outcome. Though criteria for causality in statistical studies have been researched intensely, Pearl has shown that confounding variables cannot be defined in terms of statistical notions alone; some causal assumptions are necessary [10]. In a 1965 paper, Austin Bradford Hill proposed a set of causal criteria [11]. Many working epidemiologists take these as a good place to start when considering confounding and causation.

There are various ways to modify a study design to actively exclude or control confounding variables [12]:

- Case-control studies assign confounders to both groups, cases and controls, equally. For example if somebody wanted to study the cause of myocardial infarct and thinks that the age is a probable confounding variable, each 67 years old infarct patient will be matched with a healthy 67 year old "control" person. In case-control studies, matched variables most often are the age and sex.
- Cohort studies: A degree of matching is also possible and it is often done by only admitting certain age groups or a certain sex into the study population, and thus all cohorts are comparable in regard to the possible confounding variable. For example, if age and sex are thought to be confounders, only 40–50 years old males would be involved in a cohort study that would assess the myocardial infarct risk in cohorts that either are physically active or inactive.
- Stratification: As in the example above, physical activity is thought to be a behavior that protects from myocardial infarct; and age is assumed to be a possible confounder. The data sampled is then stratified by age group – this means, the association between activity and infarct would be analyzed per each age group. If the different age groups (or age strata) yield much different risk ratios, age must be viewed as a confounding variable. There are statistical tools like Mantel-Haenszel methods that deal with stratified data.

All these methods have their drawbacks. This can be clearly seen in the following example: a 45 year old African-American from Alaska, who is an avid football player and vegetarian, working in education, suffers from a disease and is enrolled into a case-control study. Proper matching would call for a person with the same characteristics, with the sole difference of being healthy – but finding one would be an enormous task. Additionally, there is always the risk of over- and under-matching of the study population. In cohort studies, too many people can be excluded; and in stratification, single strata can get too small and thus contain only a few, non-significant number of samples [4].

An additional major problem is that confounding variables are not always known or measurable. This leads to 'residual confounding' – epidemiological jargon for incompletely controlled confounding. Hence, randomization is often the best solution since, if performed successfully on sufficiently large numbers, all confounding variables (known and unknown) will be equally distributed across all study groups.

In summary, confounding is an alternative explanation for an observed association between the exposure and outcome. Confounding is basically a mixing of effect such that the association between exposure and outcome is distorted because it is mixed with the effect of another factor that is associated with the disease. The result of confounding is to distort the true association toward the null (negative confounding) or away from the null (positive confounding). It should be re-emphasized, that a variable cannot be a confounder if it is in the causal chain or pathway. For example, moderate alcohol consumption increases serum HDL-C levels that in turn, decreases the risk of heart disease. Thus, HDL-C levels are a step in the causal chain, not a confounder that needs to be controlled [8]. Rather, this latter example is something interesting that helps us understand the disease mechanism. In contrast, because confounding factors are nuisance variables (for example, smoking is a confounder of the effect of occupational exposures (to dyes) on bladder cancer), and therefore does need to be controlled for. That is, when confounders get in the way of the relation you want to study; one wants to remove their effect. Recall that here are three ways of attenuating the effect of a confounder. The first is with the use of a case-control design, in which the confounder is matched between the cases and the controls. The second way of attenuating the effect of a confounder is mathematically, by the use of multivariate analysis. And, the third and best way to attenuate the effect of confounding is to use a randomized design; but, remember “likely to control the effect of a confounder” means just that, it’s not a guarantee.

Confounding by indication (treatment selection bias) is a bias frequently encountered in observational epidemiologic studies of drug effects. Because selection of treatments is not random and is determined by patient and physician characteristics, the observed effect is influenced by factors other than the treatment (that is the individuals at most risk are likely to be treated vs those at lesser risk), the resulting imbalance in the underlying risk profile between treated and comparison groups can generate biased results. A simple example is that subjects taking aspirin for primary prophylaxis might actually be found to have a worse outcome than the comparator group not receiving aspirin. But this latter observation might be influenced by the fact that patients taking aspirin might have had a higher disease risk burden. Once we control for disease severity and other confounders that determine who receives aspirin, we have a more accurate assessment of the relative effects of each treatment on outcome.

Confounding vs. Effect Modification

As discussed above, confounding is another explanation for apparent associations that are not due to the exposure. Also recall, that confounding is defined as an extraneous variable in a statistical or research model that affects the outcome measure, but has either not been considered or has not been controlled for during the study.

The confounding variable can then lead to a false conclusion that the outcome has a causal relationship with the exposure. Consider the example where coffee drinking is found to be associated with myocardial infarction (MI). If there is really no effect of coffee intake on MI but more coffee drinkers smoke cigarettes than non coffee drinkers, then cigarette smoking is a confounder in the apparent association of coffee drinking and MI. If one corrects for smoking, the true absence of the association of coffee drinking and MI will become apparent.

Effect modification (also referred to as interaction) is sometimes confused with confounding but with effect modification an apparent association between an exposure and outcome is "shared" with the confounder. Clinically, this can be expressed by understanding that the relationship between the exposure and outcome is different among different subgroups, or that there is a change in the magnitude of an effect according to some third variable. Referring back to the example above, let us say that coffee drinking and smoking impact on the outcome (MI). If one corrects for smoking, and there is still some impact of coffee drinking on MI, some association is imparted by cigarette smoking. In the hypothetical example above, let's say we find a RR of 5 for the association of coffee drinking and MI. When cigarette smokers are eliminated from the analysis and smoking is a confounder, the RR will be 1. In the case of effect modification where both coffee drinking and smoking equally contribute to the outcome (i.e. both smoking and coffee drinking have an equal impact on the association) the RR for each will be 2.5.

Summary

When examining the relationship between an explanatory factor and an outcome one is interested in identifying factors that may modify the factor's effect on the outcome (effect modifiers). We must also be aware of potential bias or confounding in a study because these can cause a reported association (or lack thereof) to be misleading. Bias and confounding are related to the measurement and study design. To review:

- **Bias**: A systematic error in the design, recruitment, data collection or analysis that results in a mistaken estimation of the true effect of the exposure and the outcome.
- **Confounding**: A situation in which the effect or association between an exposure and outcome is distorted by the presence of another variable. *Positive* confounding (when the observed association is biased away from the null) and *negative* confounding (when the observed association is biased toward the null) both occur.
- **Effect modification**: a variable that differentially (positively and negatively) modifies the observed effect of a risk factor on disease status. Different groups have different risk estimates when effect modification is present.

If the method used to select subjects or collect data results in an incorrect association, think bias, If an observed association is not correct because a different (lurking) variable is associated with both the potential risk factor and the outcome, but it is not a causal factor itself, think confounding; and, if an effect is real, but the magnitude of the effect is different for different groups of individuals (e.g., males vs females or blacks vs whites), think effect modification.

References

1. Cited in "Quote Me" How to add wit and wisdom to your conversation. Compiled by Edward Breslin J. Ontario: Hounslow Press; 1990. p. 44.
2. Tripod.com. Available at: http://dorakmt.tripod.com/epi/bc/html
3. Dorak MT. Bias and confounding. Available at: http://www.dorak.info/epi/bc.html
4. Systematic error. In: Wikipedia. Available at: http://en.wikipedia.org/wiki/Systematic_bias
5. Davey Smith G, Ebrahim S. Data dredging, bias, or confounding. Br Med J. 2002;325:1437–8.
6. Sampling bias. In: Wikipedia. Available at: http://en.wikipedia.org/wiki/Biased_sample
7. Sackett DL. Bias in analytic research. J Chronic Dis. 1979;32:51–63.
8. Aschengrau A, Seage GR III. Essentials of epidemiology in public health. 3rd ed. Available at: http://publichealth.jbpub.com/aschengrau/ppts/confounding.ppt10
9. Confounding. In: Wikipedia. Available at: http://en.wikipedia.org/wiki/Confounding_variable
10. Pearl J. Causality: models, reasoning, and inference. New York: Cambridge University Press; 2002.
11. Bradford Hill A. The environment and disease: association or causation? Proc R Soc Med. 1965;58:295–300. PMC1898525.
12. Hennekens C, Buring J, Mayrent SL. Epidemiology in medicine. Philadelphia: Lippincott Williams & Wilkins; 1987.

Chapter 18
It's All About Uncertainty

Stephen P. Glasser and George Howard

"Not everything that counts can be counted; and, not everything that can be counted counts" (Albert Einstein)

Abstract This chapter is aimed at providing the foundation for common sense issues that underlie why and what statistics is, so it is not a math chapter, relax! We will start with the concepts of "the universe" and a "sample", discuss the conceptual issues of estimation and hypothesis testing and put into context the question of how certain are we that a research result in the sample studied reflects what is true in the universe.

Keywords Estimation • Hypothesis testing • Statistical power • Univariate statistics • Multivariate statistics • Bayesian analysis

It is surprising that as a society we accept poor math skills. Even if one is not an active researcher, one has to understand statistics to read the literature. Fortunately, most of statistics are common sense. This chapter is aimed at providing the foundation for common sense issues that underlie why and what statistics is, so it is not a math chapter, relax! As one popular cartoon portrayed " no one will enter heaven without answering the question related to a train leaving the station at 12 noon traveling in one direction at 100 mph and another train is traveling towards it leaving at 1 PM and traveling at 60 mph when will they meet?"

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

G. Howard, Ph.D.
Department of Biostatistics, School of Public Health, University of Alabama at Birmingham,
Birmingham, AL, USA

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_18,

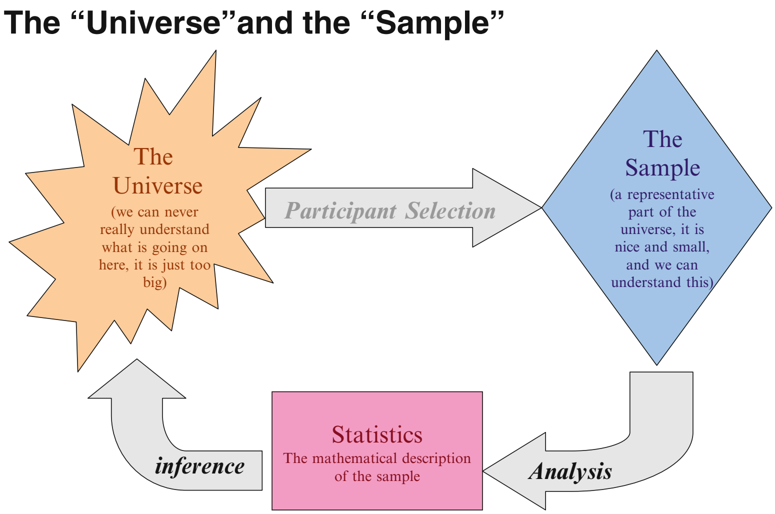

The "Universe"and the "Sample"
The Universe
(we can never really understand what is going on here, it is just too big)
Participant Selection
The Sample
(a representative part of the universe, it is nice and small, and we can understand this)
Analysis
Statistics
The mathematical description of the sample
inference

Fig. 18.1 What is the purpose of statistics? Characterizing the universe from a sample

Let's start with the concepts of "the universe" and of a "sample". The "universe" is that group of people (or things) that we really want to know about … it is what we are really trying to describe. For example, for a study we might be interested in the blood pressure for British white males – then the "universe" is every white man in the Great Britain. The trouble is that the "universe" is simply too big for research purposes, we cannot begin to measure everybody in Great Britain, so we select a representative part of the universe – which is our study sample. Since the sample is much smaller than the universe we have the ability to measure things on everybody in the sample and analyze relationships between factors in the sample. If the sample is really representative of the universe, and we understand what is going on in that sample, we gain an *inferential* understanding of what is happening in the universe. The critical concept is that we perform our analysis on the sample (which is not what we really want to describe) and infer that we understand what is going on in the universe (which is our real goal) (as an aside, when the entire universe is measured it is called performing a *census* and we all know even that has its problems). There are, however, advantages of measuring everyone if we could. For example, if we could measure everyone, we will get the correct answer – there is almost no uncertainty when everyone is measured; and, one will not need a statistician – because the main job of a statistician is to deal with the uncertainty involved in making inferences from a sample. However, since measuring everyone is impractical (impossible?), and very expensive, for practical reasons one is forced to use an inferential approach, which if done correctly, one can *almost* be certain to get *nearly* the correct answer. The entire field of statistics deals with this uncertainty, specifically to help define or quantify "almost" and "nearly" when making an inference (Fig. 18.1). The characteristic that defines any statistical approach is how it deals with uncertainty. The traditional approach to dealing with uncertainty is the Frequentist approach, which assumes the existence of "parameters" which represent the "correct answer" in the universe. For example, is there a true height of British white males. The Frequentist approach then takes a sample of British while males, measures how tall they are, and then "guesses" the value of the parameter. The trick of statistics is that this guess comes with a measure of the uncertainty in the guess … that is, the guess may be that British white males are 5 ft, 11 in. tall, but this guess comes with a statement that we are pretty certain (say 95 %) that the true value is somewhere between 5 ft, 9 in. and 6 ft, 1 in. A Bayesian approach can use previous data to develop a prior distribution of potential heights of white male British, and updates this data with that collected to develop a posterior distribution (this is akin to the discussion in Chap. 14 that addresses pre and post-test probability). We will discuss the Bayesian approach later in this chapter. However, the focus of this chapter is the Frequentist approach.

There are two kinds of inferential activities statisticians perform – estimation and hypothesis testing, each described below.

Conceptual Issues in Estimation

Estimation is simply the process of producing a very educated guess for the value of some parameter ("truth") in the universe. In statistics, as in guessing in other fields, the key is to understand how close the estimate is to the true value. Conceptually, *parameters* (such as an average BP of men in the US) exist in the *universe* and do not change, but we cannot know them without measuring everyone. The natural question would then be "*how good is our guess*;" and, for this we need to have some measure of the reliability of our estimate or guess.

If we select two people out of the universe, one would not expect them to have the same exact measurement (i.e. for example, we would not expect them to have the identical blood pressure). People in a population have a dispersion of outcomes that is characterized by the standard deviation. We might recall from standardized testing for college and graduate programs that about 95 % of the people are within about 2 standard deviations of the average value. That is, getting people who are more than two standard deviations away from the mean will not happen very often (in fact, less than 5 % of the time).

Returning to the example mentioned above, suppose we are interested in estimating (guessing) the mean blood pressure of white men in Great Britain. How much variation (uncertainty) can we reasonably expect between two estimates of the mean blood pressure? To answer this, consider that the correct answer exists in the universe, but the estimate from a sample will likely be somewhat different from that true value. In addition, a different sample would likely give a result that is both different from the "true" value and different from the first estimate. If one repeats the experiment in a large number of samples, the different estimates that would be produced from the repeated experiments would have a standard deviation. This standard deviation of estimates from repeated estimates has a special name – the *standard error of the estimate*. The standard error is nothing more than the standard deviation of the estimate, if the same experiment was repeated a large number of times. That is, if one repeats an experiment 100 times, (i.e. obtain100 different samples of white men, and each time calculate a mean blood pressure), just as we would not expect individual people to have the same blood pressure, we would not expect these samples to have the same mean blood pressure. The standard deviation of means is called the standard error of the mean. The real trick of the field of statistics is to provide an estimate of the standard error of a parameter when the experiment is only performed a single time. That is, if a single sample is drawn from the universe, on the basis of that single sample is it possible to say how much you would expect the mean of future samples to differ from that obtained in this first sample (if one thinks about it … that is quite a trick).

As mentioned above, we are all familiar with the fact that 95 % of people are within two standard deviations of the mean (again, think about the standardized tests we have all taken). It turns out that 95 % of the estimates are also within two standard deviations (except we call it two standard errors) of the true mean. This observation is the basis for "confidence intervals" and this can be used to characterize the uncertainty of the estimation. The calculation of a confidence interval is nothing more than a reflection of the same concept that 95 % of the people (estimates) are within about two standard deviations (standard errors) of the mean. The use of confidence intervals permits a more refined assessment of the uncertainty of the guess, and is a range of values calculated from the results of a study, within which the true value lies; the width of the interval reflecting random error. The width of the confidence limit differs slightly from the two standard errors, (due to adjustment for the uncertainty from sampling), and the width is also a function of sample size (a larger sample size reduces the uncertainty). Also, the most common interpretation of a confidence interval is that "I am 95 % sure that the real parameter is within this range" is technically incorrect, albeit not that incorrect. The correct interpretation is much less intuitive (and therefore is not as frequently used) – that if an experiment were repeated a large number of times, and 95 % confidence limits were calculated each time using similar approaches, then 95 % of the time these confidence limits would include the true parameter. We are all accustomed to hearing about confidence limits, since confidence intervals are what pollsters mean when they talk about the "margin of error" of their poll.

To review, estimation is an educated guess of a parameter, and every estimate (not only estimated means, but also estimated proportions, slopes, and measures of risk) has a standard error. The 95 % confidence limits depict the range that we can "reasonably" expect the true parameter to be within (approximately ±2 SE). For example, if the mean SBP is estimated to be 117 and the standard error is 1.4, then we are "pretty sure" the true mean SBP is between 114.2 and 119.8 (the slightly incorrect interpretation of the 95 % confidence limit is "I am 95 % sure that the real parameter is between these numbers").

Table 18.1 Types of estimates of the measure of association

		Full professor by age 40		
		Yes	No	Total
Attended course	Yes	20	11	31
	No	8	12	20
	Total	28	23	

Proportion that achieved goal among those reading this book 20/31 = 65 %
Proportion that achieved goal among those not reading this book 8/20 = 40 %
Difference between the two proportions .65 – .40 = .25 or a 25 % increase in success
The RR of achieving the goal 0.65/0.40 = 1.61 (about a 61 % increase)
The OR of achieving the goal the odds of achieving the goal by reading the book 20/11 = 1.81; achieving the goal in those not reading the book 8/12 = 0.67; the odds ratio is 1.81/0.67 = 2.73

Studies frequently focus on the association between an "exposure" (treatment) and an "outcome". In that case, parameter(s) that describe the strength of the association between the exposure and the outcome are of particular interest. Some examples are:

- The difference in cancer recurrence at a given time, between those receiving a new versus a standard treatment
- The reduction in average SBP associated with increased dosage of an antihypertensive drug
- The differences in the likelihood of being a full professor before age 40 in those who read this book versus those who do not

Let's say we have a sample of 51 University of Alabama at Birmingham students some of whom have read an early draft of this book years ago. We followed each of these students to establish their academic success, as measured by whether they made the rank of full professor by age 40. The, resulting data is portrayed in Table 18.1. From a review of Table 18.1 what types of estimates of the measure of association can we make from this sample? We can:

1. Calculate the *absolute difference* in those achieving the goal.
   (a) Calculating the proportion that achieved the goal among those reading the book (20/31 = 0.65 or 65 %)
   (b) Calculating the proportion that achieved the goal among those not reading the book (8/20 = .40 or 40 %)
   (c) By calculating the difference in these two proportions (0.65 – 0.40 = 0.25), we can demonstrate a 25 % increase in the likelihood of academic success by this measure.

Or…

2. We can calculate the *relative risk (RR)* of achieving the goal
   (a) By, calculating the proportion that achieved the goal among those reading the book (20/31 = 0.65 or 65 %)
   (b) By, calculating the proportion that achieved the goal among those not reading the book (8/20 = .40 or 40 %)

Table 18.2 Major points regarding estimation

Estimates from samples are only educated guesses of the truth
Every estimate has a standard error, which is a measure of the variation in the estimates
If you were to repeat a study, one should no: expect to get the same answer
When you have two estimates, you can conclude:
It is almost certain that neither is correct
However, in a well-designed experiment
The guesses should be "close" to "correct"
Statistics can help us understand how far our guesses are likely to be from the truth…
And how far they would be from other guesses

(c) And then calculating the ratio of these two proportions (RR is 0.65/0.40 = 1.61) – or there is a 61 % increase in the likelihood of making full professor among those reading the book.

Or…

3. We can calculate the *odds ratio (OR)* of achieving this goal:

(a) By calculating the odds (the "odds" is the chance of something happening divided by the chance of it not happening) of achieving the goal among those reading the book (20/11 = 1.81)
(b) By calculating the odds of achieving the goal among those not reading the book (8/12 = 0.67)
(c) And then, calculating the ratio of these two odds (OR is 1.81/0.67 = 2.73) – or there is a 2.73 times greater odds of making full professor among those reading the book.

The point of this example is to demonstrate that there are different estimates that can reasonably be produced from the very same data. Each of these approaches is correct, but they give extremely different impressions of what is occurring in the study (that is, is there a 25 % increase, a 65 % increase or a 173 % increase?). In estimation, therefore, great care should be taken to make sure that there is a deep understanding of what is being estimated.To review the major points about estimation (Table 18.2):

- Estimates from samples are only educated guesses of the truth (of the parameter)
- Every estimate has a standard error, which is a measure of the variation in the estimates. When standard errors are not provided, care should be taken in the interpretation of the estimates – they are guesses without an assessment of the quality of the guess (by the way, note that standard errors were not provided for the guesses

made from Table 18.1 of the difference, the relative risk, or the odds ratio of the chance of making full professor).

- If you were to repeat a study, one should not expect to get the same answer (just like if one sampled people from a population, one should not expect them to have the same blood pressure amongst individuals in that sample).
- When you have two estimates, you can conclude:
  - It is almost certain that neither is correct
  - However, in a well-designed experiment
    - The guesses should be "close" to "correct"
    - Statistics can help us understand how far our guesses are likely to be from the truth, and how far they would be from other guesses (were they made).

Conceptual Issues in Hypothesis Testing

The other activity performed by statisticians is hypothesis testing, which is simply making a yes/no decision regarding some parameter in the universe. In statistics, as in other decision-making areas, the key to decision-making is to understand what kind of errors can be made; and, what the chances are of making an incorrect decision. The basis of hypothesis testing is to assume that whatever you are trying to prove is not true –i.e. that there is no relationship (or technically, that the null hypothesis H_o is supported). To test the hypothesis of no difference, one collects data (on a sample), and calculates some "test statistic" that is a function of that data. In general, if the null hypothesis is true, then the test statistic will tend to be "small;" however, if the null hypothesis is incorrect the test statistic is likely to be "big." One would then calculate the chance that a test statistic as big (or bigger) as we observed would occur under the assumption of no relationship (this is termed the p-value!). That is, if the observed data is unlikely under the null, then we either have a strange sample, or the null hypothesis of no difference is wrong and should be rejected. To return to Table 18.1, let's ask the question "how can one calculate the chance of getting data this different for those who did versus those who did not read this book, under the assumption that reading the book has no impact?" The test statistic is then calculated to assess whether there is evidence to reject the hypothesis that the book is of no value. Specifically, the test statistic used is the Chi-square (χ^2), the details of which are unimportant in this conceptual discussion – but the test statistic value for this particular table is 2.95. Now the question becomes is 2.95 "large" (providing evidence that the null hypothesis of no difference is not likely) or "small" (failing to provide such evidence). It can be shown that in cases like the one considered here, that if there is really no association between reading the book and the outcome, that only 5 % of the time is the value of the test statistic larger than 3.84 (this, therefore, becomes the definition of "large"). Since 2.95 is less than 3.84, this is not a "large" test statistic; and, therefore, there is not evidence to support that the null hypothesis is wrong (i.e. that reading the book has no impact is wrong – however, one cannot use these hypothetical data to prove that you are currently otherwise spending your time wisely). We acknowledge and regret that this double-negative statement must be made, i.e. "there is not evidence that the null hypothesis is wrong". This is because, one does not "accept" the null hypothesis of no effect, one just does not reject it. This is a small, but critically important concept in hypothesis testing – that a "negative" test (as was true in the above example) does not prove the null hypothesis, it only fails to support the alternative. On the other hand, if the test statistic had been bigger than 3.84, then we would have rejected the null hypothesis of no difference and accepted the alternative hypothesis of an effect (i.e. that reading this book does improve one's chances of early academic advancement-obviously the correct answer).

P Value

The "*p-value*" is the chance that the test statistic from the sample could have happened under the null hypothesis. What constitutes a situation where it is "unlikely" for the data to have come from the null, that is, how much evidence are we going to require before one "rejects" the null? The standard is that if the data has less than a 5 % chance ($p<0.05$) of happening by chance alone, then the observation is considered "unlikely". One should realize that this p value (0.05) is an arbitrary number, and many argue that too much weight is given to the p-value. None-the-less, the p-value being less than or greater than 0.05 is inculcated in most scientific work. However, consider the example of different investigators performing an identical experiment and one gets $p=0.053$, whereas the other gets $p=0.049$. Should one really come to different conclusions? In one case there is a 5.3 % chance of getting data as observed under the null hypothesis, and in the other there is a 4.9 % chance. If one accepts the 0.05 threshold as "gospel," then these two very similar results appear to be discordant. Many people do, in fact, adhere to the position that they are "different" and are discordant, while others feel that they are confirmatory. To make things even more complex, one could argue that the interpretation of the p value may depend on the context of the problem (that is, should one always require the same level of evidence?). See Table 3.12 for a list of some common p-value misinterpretations.

Aside from the arguments above, there are a number of ways to "mess up" the p value. One certain way is to not follow the steps in hypothesis testing, one surprising, but not uncommon way to mess things up. Consider the following steps one researcher took: after looking at the data the investigator created a hypothesis, tested that hypothesis, and obtained a p-value; that is, the hypothesis was created from the data (see discussion of subgroup and post-hoc analysis). Forming a hypothesis from data already collected is frequently referred to as "data dredging" (a polite term for the same activity is "exploratory data analysis"). Another way of messing up the p value is to look at the data multiple times during the course of an experiment. If one looks at the data once, the chance of a spurious finding is 0.05; but with multiple "peeks", the chance of spurious findings increase significantly (Fig. 18.2). For example, if one "peeks" at the data five times during the course of one's experiment, the chance of a spurious finding

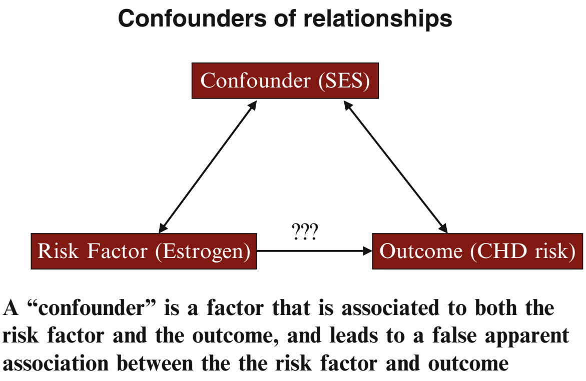

Confounders of relationships
Confounder (SES)
???
Risk Factor (Estrogen)
Outcome (CHD risk)
A "confounder" is a factor that is associated to both the risk factor and the outcome, and leads to a false apparent association between the the risk factor and outcome

Fig. 18.2 Depicts an example of trying to prove an association of estrogen and CHD (indicated by the *question marks*) but that socioeconomic status (SES) is a factor that influences the use of estrogen and also affects CHD risk separate from estrogen. As such, SES is a confounder for the relationship between estrogen and CHD risk

increases to almost 20 % (i.e. we went from 1 chance in 20 to about a 4 in 20 chance of a spurious finding). What do we mean by peeking at the data? This frequently occurs from: interim examinations of study results; looking at multiple outcome measures; analyzing multiple predictor variables; or, performing subgroup analyses. Of course, all of these can be legitimate; it just requires planning (that is pre-planning).

Regarding subgroup analysis, It is not uncommon that after trial completion, and while reviewing the data one discovers a previously unsuspected relationship (i.e. a post-hoc observation). Because this relationship was not an *a priori* hypothesis, the interpretation of the p value is no longer reliable. Does that mean that one should ignore the relationship and not report it in one's manuscript? Of course not, it is just that one should be honest about the conditions of the discovery of the observation. What should be said in the paper is something similar to:

> In exploratory analysis, we noted an association between X and Y. While the nominal p-value of assessing the strength of this association is 0.001, because of the exploratory nature of the analysis we encourage caution in the interpretation of this p-value and encourage replication of the finding.

This is a "proper" and honest statement that might have been translated from:

> We were poking around in our data we found something that is really neat. We want to be on record as the first to report this. We sure do hope that you other guys see this in your data too.

Type Error I, Type II Error, and Power

To this point, we have been focusing on a specific type of error – one where there really is no difference (null hypothesis is true) between the groups, but we are concerned about falsely saying there is a difference. This would be akin to a false positive result and this is termed a "Type I Error." Type II errors occur if one says there

Table 18.3 A depiction of Type I and Type II error

	Null hypothesis (No difference)	Alternative hypothesis (There is a difference)
Test conclusion: No evidence of difference	You win! Correct decision	You lose! Incorrect decision B = Type II error
Test conclusion: There is a difference	You lose! Incorrect decision α = Type I error	You win! Correct decision $1 - \beta$ = Power

is not evidence of a difference when a difference does indeed exist; and this is akin to a false negative result. To recap, recall that one initially approaches hypothesis testing with the statement that there was no difference (the null hypothesis is true), one then calculated the chance that a difference as big as the one you observed in the data was due to chance alone, and if you reject that hypothesis ($P < .05$), you say there really is a difference, then the p value gives you the chance that you are wrong (i.e. $p < .05$ means there is less than 1 chance in 20 that you are wrong and 19 chances out of 20 that you are right- i.e., that there really is a difference). Table 18.1 portrays all the possibilities in a 2×2 table (Table 18.3).

Statistical Power (Also See Chap. 15)

Statistical power, is the probability that given that the null hypothesis is false (i.e. that there really is a difference) that we will see that difference in our experiment. Power is influenced by:

- The significance level (α): if we require more evidence to declare a difference (i.e. a lower p value –say $p < .01$) , it will be harder to get, and the sample size will have to be larger, as this determination will allow one to provide for greater (or less) precision (i.e. see smaller differences);
- the true difference: this is from the null hypothesis (i.e. big differences are easier to see than small differences)
- The other parameter values related to "noise" in the experiment. For example, if the standard deviation (δ) of measurements within the groups is larger (i.e., there is more "noise" in the study) then it will be harder to see the differences that exist between groups
- The sample size (n). It is not wrong to think of sample size as "buying" power. The only reason that a study is done with 200 rather than 100 people is to buy the additional power.

To review, some major conceptual points about hypothesis testing are:

- Hypothesis testing is making a yes/no decision
- The order of steps in statistical testing is important (the most important thing is to state the hypothesis before seeing the data)

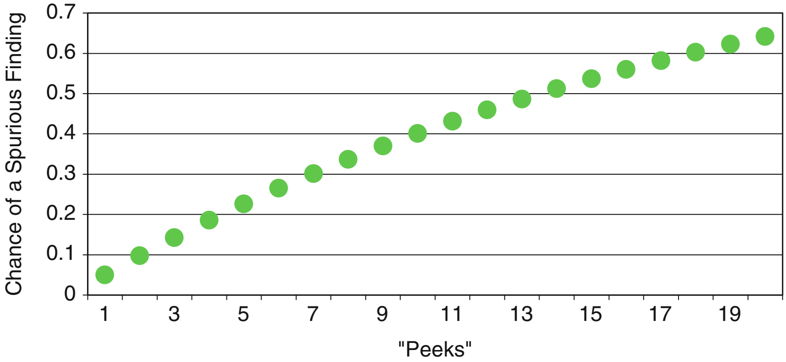


Fig. 18.3 The chance of spurious findings related to the number of times the data is analyzed during the course of a trial

- There are many ways to make a mistake, including
  - Saying there is a difference when there is not one
    - By design, the α level gives the chance of a Type I error
    - The p-value is the chance in the specific study
  - Saying there is not a difference when there is one
    - By design, the β level gives the chance of a type II error, with 1- β being the "power" of the experiment
    - Power is the chance of seeing a difference when one truly exists
- P-values should be interpreted in the context of the study
- Adjustments should be made for multiple "peeks" (or interpretations should be made more carefully if there are multiple "peeks") – See Fig. 18.3

Univariate and Multivariate (Multivariable) Statistics

To understand these analyses one must have an understanding of confounders (also see Chaps. 3 and 17). A confounder is a factor that is associated with both the exposure (say a risk factor) and the outcome; and, leads to a false apparent association between the two. Let's use, as an example, the past observational data on the beneficial association of hormone replacement therapy and beta-carotene on atherosclerosis, MI and stroke risk. When RCTs were performed, these associations not only disappeared, but there was a suggestion that some of these exposures were potentially harmful. Confounders are one of the major limitations of observational studies, (recall that for RCTs, randomization equalizes known and unknown confounders between the intervention and control groups so they are not a factor in the observed associations). In observational studies, however, it is necessary to "fix" the effect of confounders on the association one is trying to evaluate. In observational studies there are two basic ways of "fixing" confounders: (1) match the interventional and control groups for known confounders, at the start of the study, or (2) to adjust for potential confounders during data analysis (See discussion of Propensity Scoring in Chap. 3). One should note that either of these approaches can only "fix" ***known*** confounders, which is unlike randomization that also "fixes" any ***unknown*** confounders (this being one of the major reasons that RCTs result in the highest level of scientific evidence). Remember too, that for something to be a confounder it must be associated with both the exposure and the outcome. In a case-control study, for example, one matches the cases and controls (for example by matching for age and race) so that there can be no association between those confounders (age and race) and the outcome (i.e., the cases and controls have the same distribution of race and age – because they were made to).

A way to mathematically adjust for confounders is multivariate analysis. That is, in case-control, cross-sectional, or cohort studies, differences in confounders between those with and without the "exposure" can be made to be equal by mathematical adjustment. Covarying for confounders is the main reason for multivariate statistics. The interpretation of the exposure variable in a multivariate model is "the impact of a change in the exposure variable at a fixed level of the confounding variable(s)." Saying that the association of the predictor and the outcome "is at a fixed level of the confounding variable" is the same as saying that there is not an association between the exposure and the confounding variable (really, that the relationship has been "accounted for").

Again however, many things can "go wrong" in multivariate analysis. As already mentioned, one must know about the confounders in order to adjust or match for them. In addition, one must be able to appropriately measure confounders (take SES for example, since there is much argument as to what components should make up this variable the full effect of SES may be difficult to account for in the analysis). Not only can one not quantify parts of a confounder, a confounder can never be perfectly measured and as a result confounders cannot be perfectly accounted for. Also, even when a potential confounder is identified, the more measurement error there is in the confounder, the more likely that "residual confounding" can still occur.

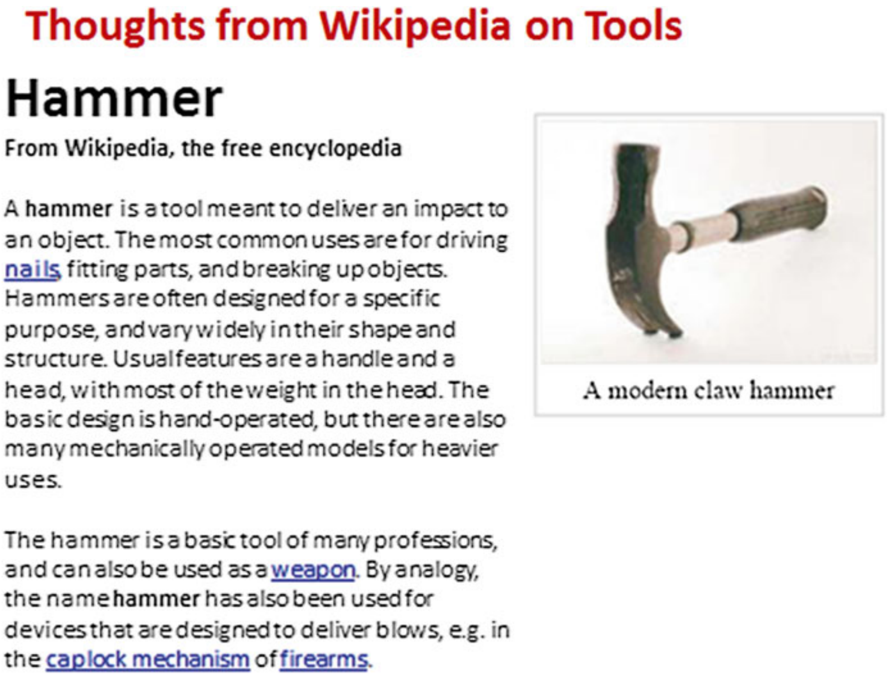
Bayesian Analysis

One of the many confusing statistical concepts for the non-statistician is the argument over which approach- Frequentist or Bayesian-is preferable. With the Frequentist approach (this has become the traditional approach for clinical trials) an assumption is made that the difference between treatment groups is unknown and the parameter is fixed (for example, the mean SBP of all British citizens is a fixed number). With the Bayesian approach (some argue becoming a much more common approach in the future) parameters are assumed to be a distribution of potential differences between treatment groups and that there is information existent about these differences before you do the proposed trial. This latter idea defines one of the major strengths of the Bayesian approach- that is that one can use prior information (prior distribution) known from other studies before one conducts their trial, and this can be "added to the information gained in the current trial (posterior distribution) with the potential benefit of a reduced sample size necessary to show a difference (with the Frequentist approach one starts statistically with a 'clean slate')". Howard et al argues for the Frequentist approach by noting that "*we have a difficult time agreeing what we know*"-that is the choice of studies guiding the prior knowledge is largely subjective [1]. Frequentists also argue that if there is little prior knowledge, there would be no meaningful reduction in sample size, while substantial prior knowledge brings into play the ethical need to do a new study. Finally, Frequentists argue, that there are at least two reasons why previous studies might provide incorrect information (sampling variation which can be adjusted for, and bias which cannot) and the inclusion of these in the prior distribution then adversely affects the posterior distribution [1]. Berry argues that the Bayesian approach is optimal because it is "tailored to the learning approach", that is as information is accrued one "updates what one knows"; and, that this flexibility makes it ideal for clinical research [2]. Howard argues that rather than being one or the other, one should be an opportunist. Howard makes this latter point using thoughts from Wikipedia on Tools: as an opportunist, one should take advantage of the optimal tools for the problem at hand; only a fool would use the same tool for all problems!

- When one needs to drive a nail, a hammer is a neat tool
- When one needs to cut a board, a saw is awfully handy

Thoughts from Wikipedia on Tools (Figs. 18.4 and 18.5)

The question then is: *Does a Bayesian approach provide a meaningful advantage over the traditional approach particularly if we focus on the Phase III randomized clinical trial?* Bayesian Analysis really pays off if there is substantial "prior" information that can be used to powerfully inform the posterior information. Some argue that Bayesian analysis allows adaptive designs to be implemented, and this is true! But adaptive designs can also be implemented under a Frequentist approach. Some argue that Bayesian analysis "answers the logical question" … actually it can be easily argued that both approaches can be used to answer the logical questions. Bayesian approaches have the advantage of optimally using the information available from previous studies to inform the Phase III trials. This allows for the "smooth transition from Phase II to Phase III trials". And, who would not use all the information that is available? As briefly discussed above, the problem is that we may not agree on "what we know", and there are at least three issues in this regard: *we have to agree on what we know; and, we have to know just the right amount; and, we may be systematically wrong about both.*



Thoughts from Wikipedia on Tools

Hammer

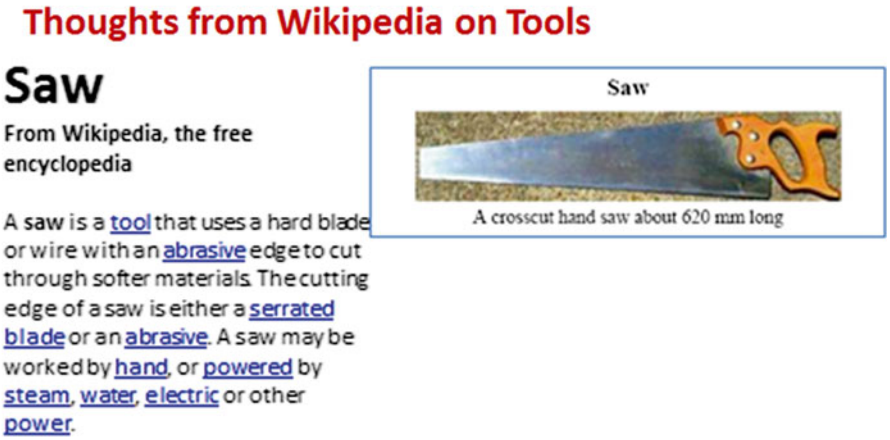
From Wikipedia, the free encyclopedia

A **hammer** is a tool meant to deliver an impact to an object. The most common uses are for driving nails, fitting parts, and breaking up objects. Hammers are often designed for a specific purpose, and vary widely in their shape and structure. Usual features are a handle and a head, with most of the weight in the head. The basic design is hand-operated, but there are also many mechanically operated models for heavier uses.

The hammer is a basic tool of many professions, and can also be used as a weapon. By analogy, the name **hammer** has also been used for devices that are designed to deliver blows, e.g. in the caplock mechanism of firearms.

A modern claw hammer

Fig. 18.4 Contrasting frequentist and Bayesian statistical approaches using a hammer and saw example



Thoughts from Wikipedia on Tools

Saw

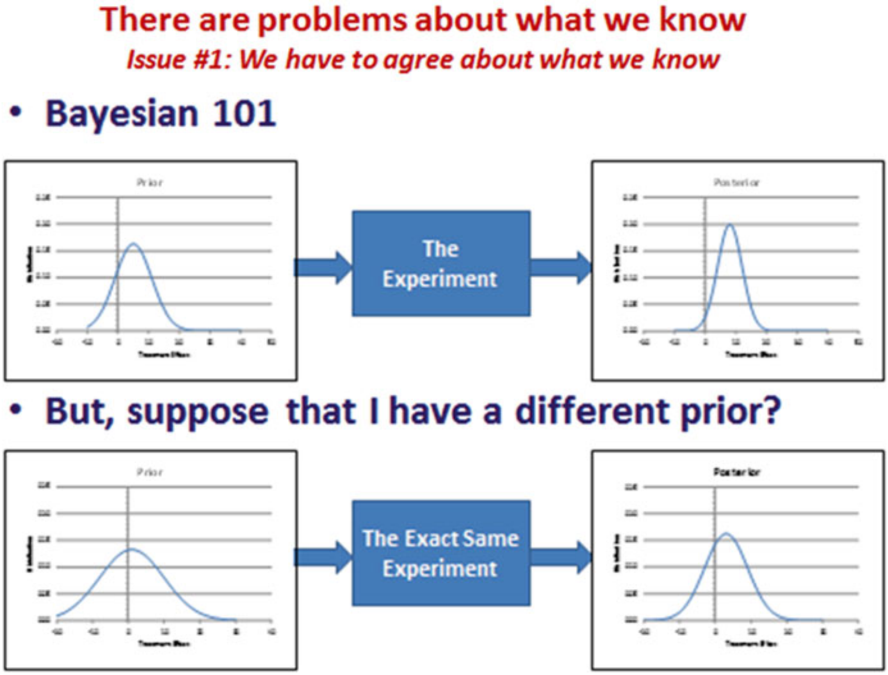
From Wikipedia, the free encyclopedia

A **saw** is a tool that uses a hard blade or wire with an abrasive edge to cut through softer materials. The cutting edge of a saw is either a serrated blade or an abrasive. A saw may be worked by hand, or powered by steam, water, electric or other power.

Saw

A crosscut hand saw about 620 mm long

Fig. 18.5 Contrasting frequentist and Bayesian statistical approaches using a hammer and saw example



There are problems about what we know

Issue #1: We have to agree about what we know

- Bayesian 101

The Experiment

- But, suppose that I have a different prior?

The Exact Same Experiment

Fig. 18.6 An example demonstrating the Bayesian approach

Issue #1: We Have to Agree About What We Know

Consider that in academics, we write papers, we care for patients, and we teach students. What we do poorly is agree about what we know. There are many examples of large clinical trials in which experts disagree about what the results mean (Fig. 18.6).

Issue #2: We Have to Know Just the Right Amount

Remember, we only truly gain from Bayesian analysis if we know something, so if we know nothing … we gain virtually nothing. Also remember that randomization is based on equipoise, so if we know too much, we are not at equipoise. How would an Informed Consent statement sound if we said w*e are only 80 % sure that you will benefit from the new treatment; however, to remove our uncertainty, we would like to assign you at random to treatment*. Where does one draw the line … if we are 60 %/40 % … how about 70 %/30 %?

Issue #3: We May Be (Systematically) Wrong About What We Know

We can unrealistically pretend that the current study is the only study, but let's consider that there are 100's of compounds introduced to assess efficacy and of these only a few will make it through Phase II to a Phase III trial. The reason why a compound advances is that it is promising, and there are two reasons for a compound to be promising:

- It works!!!
- Through a random process, it (inappropriately) appears to work

Thus, conditional on making it to a Phase III trial, the Phase II results are biased. But, If Phase II results are biased, then incorporating that information will also bias the Phase III results. Also, one might ask: *Do we really want to use that prior information anyway*? The knee-jerk is … we need to use all the information available to make a decision. But, suppose we are getting ready to do a pivotal trial, does it make sense to "*wipe the board clean*" and have a true independent test of the compound? Is it a bad decision to have an independent assessment of the new treatment? On the other hand, do we really understand the prior information?

- Could some of the previous scientists have "had an agenda"?
- Is our understanding of the information really the true status of the information

Thus, should the argument about the Frequentist Approach vs the "Bayesian Approach" really be, we are just someone who tries to understand the best tools for a problem.

Selection of Statistical Tools (Or Why Are There So Many Statistical Tests?)

Each research problem can be characterized by the type and function of the variable and whether one is doing single or repeated assessments of the data. These are the characteristics of an experiment that determine the statistical tool used in the study. The first characteristic that influences the choice of which statistical "tool" to use, is the *data type*. Data types are categorical, ordinal or continuous. Categorical data (also called nominal or dichotomous if one is evaluating only two groups), is data that are in categories i.e. neither distance nor direction is defined e.g. gender (male/female), ethnicity (AA, NHW, Asian), or outcome (dead/alive), hypertension status (hypertensive, normotensive). Ordinal data, is data that are in categories and direction but not distance, good/better/best; normotensive, borderline hypertension, hypertensive. With continuous (also called interval) data, both distance and direction are defined e.g. age or systolic blood pressure.

Data function is another characteristic to consider. With data function, we are dealing with whether the data is the dependent or independent variable. The dependent

Table 18.4 The statisticians "Toolbox"

		Type of independent data					
		Categorical				Continuous	
		Two samples		Multiple samples			
Type of dependent data	One sample (focus usually on estimation)	Independent	Matched	Independent	Repeated measures	Single	Multiple
Categorical (dichotomous)	1 Estimate proportion (and confidence limits)	2 Chi-square test	3 McNemar test	4 Chi-square test	5 Generalized Estimating Equations (GEE)	6 Logistic regression	7 Logistic regression
Continuous	8 Estimate mean (and confidence limit)	9 Independent t-test	10 Paired t-test	11 Analysis of variance	12 Multivariate analysis of variance	13 Simple linear regression and correlation coefficient	14 Multiple regression
Right censored (survival)	15 Kaplan Meier Survival	16 Kaplan Meier Survival for both curves, with tests of difference by Wilcoxon or log-rank test	17 Very unusual	18 Kaplan-Meier Survival for each group, with tests by generalized Wilcoxon or Generalized Log Rank	19 Very unusual	20 Proportional Hazards analysis	21 Proportional Hazards analysis

variable is the outcome in the analysis and the independent variable is the exposure (predictor or risk factor).

Finally, one needs to address whether single or repeated assessments are being performed. That is, a single assessment is a variable that is measured once on each study participant (for example baseline blood pressure measured on two different participants); while repeated measures (if there are two measures, it is also called "paired measures") are measurements that are repeated multiple times (frequently at different times), for example, repeated measures on the same participant at baseline and then 5 years later, or blood pressures of siblings in a genetic study (in this latter case the study is of families not people, and there are two measures on the same family). Why do there have to be so many approaches to these questions? Just as a carpenter needs a saw and a hammer for different tasks, a statistician needs different types of analysis tools from their "tool box" (Table 18.4).

References

1. Howard G, Coffey C, Cutter G. Is Bayesian analysis ready for use in Phase III randomized clinical trials? Beware the sound of sirens. Stroke. 2005;36:1622–3.
2. Berry D. Is the Bayesian approach ready for prime time? Yes! Stroke. 2005;26:1621–2.

Chapter 19
Grant and Manuscript Writing

Donna K. Arnett and Stephen P. Glasser

Abstract Perhaps nothing is more important to a new investigator than how to properly prepare a grant to request funding for clinical research or how to write a manuscript for publication. In this chapter we will review the basic elements for successful grant and manuscript writing, discuss advantages and disadvantages of K versus R applications for National Institutes of Health (NIH) funding, illustrate the "fundamentals" for each section for a standard NIH R-series application, and describe the key components necessary to transition to a successful NIH research career.

Keywords Research grant structure • Writing a manuscript • Journal guideline statements • CONSORT • Conflicts of interest • Coercive citations • Open access journals

Basic Tenets of Grant Writing

The three fundamental principles involved in the successful preparation of an NIH grant are to understand the mission of the particular NIH branch from which you wish to secure funding, to know the peer review process, and to build the best team possible to accomplish the work proposed. It is very important, particularly to new investigators, to secure collaborators for areas in which you lack

D.K. Arnett, Ph.D., MPH
Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_19,

experience and training. While this often proves to be challenging for the new investigator since it is difficult to secure the attention of busy senior investigators, it is a critical step toward securing funding for the work you propose. Finally, grant writing like any skill, can only be optimized by doing it repeatedly. You can read all about the physics of learning to ride a bicycle, but until one does it repetitively, one will not be good at it. The same is true with respect to grant writing: writing, editing, and re-writing of the grant should occur on a regular basis.

Having all the tools described above in your "toolbox", however, will not necessarily lead to a successful grant. The ideas must be presented, or "marketed" in such a way as to show the review team the importance of the proposed work as well as its innovative elements. The grant proposal must be presented in an attractive way and the information placed where reviewers expect to find it. Complex writing styles are also ill advised for grants. It is important to use clear and simple sentence structures, and to avoid complicated words. Also avoid the temptation to use abbreviations to save space since many abbreviations, or unusual abbreviations, make a grant difficult to read. Instead, use a reviewer friendly approach where the formatting is simple and the font is readable. Organize and use subheadings effectively (e.g., like a blueprint to the application), and use topic sentences for each section that build the "story" of your grant in a logical and sequential way. Use spell-checking programs before submission, and also, ask a colleague to read through the final draft before submission. Most importantly, be consistent in specific aims and format throughout the application.

Very importantly, the proposal must convince evaluators that the problem you are addressing is critical and significant, and that the team can deliver. Also, it is important to recognize that when investigators possess knowledge about a subject, it is hard for them to imagine what it is like not to know, and this is referred to as "the curse of expertise" [1]. This "curse" prevents effective communication. VanEkelenberg also points to the "chain of reasoning" to refer to the importance of a "roadmap that guides the reader through the proposal". He also has developed a table (Table 19.1) that provides a model for the chain of reasoning.

Table 19.1 The **PROSANA** model for developing the "chain of reasoning"

Step	Guide word	Explanation
1.	Problem	Carefully describe the perceived problem
2.	Root causes	Describe the underlying causes in statements
3.	fOcus	Narrow the problem by focusing on the causes addressed by the proposal
4.	Solutions	Briefly mention potential solutions making clear that the writer is aware of alternative approaches
5.	Approach	Narrow the approach to the chosen solution for the proposal
6.	Novelty	Describe the associated novelty either in the approach, technology, etc.
7.	Arguments	List the main arguments that explain/support the logic for the proposed solution

The Blueprint of a Research Grant

For the scientist, the most important content of the NIH grant for which the proponent is fully responsible consists of the:

Abstract
Budget for initial period
Budget for 5-year period
Introduction (Revised or Supplemental applications)
Research Plan, which includes:

- Specific Aims
- Background and Significance
- Preliminary Studies/Progress Report
- Research Design and Methods
- Use of Human Subjects
- Use of Vertebrate Animals
- Literature Cited
- Data Sharing Plan

There are many administrative forms that also must be included from your agency (such as the face page and the checklist, to name a few), but the items described above are where you will spend the majority of your time. It is important to carefully read the instructions, and also to check with your agency's grants and contracts officer to resolve any questions ***early*** in the process of preparing your application.

Writing the Research Grant

In writing the research grant, start with strengths by clearly articulating the problem you will address and how it relates to the present state of knowledge. Find the gap in knowledge and show how your study will fill that gap and move the field closer to the desired state of knowledge. Pick the "right" question, knowing that the question should have potential to get society closer to an important scientific answer while at the same time knowing that there are many, more questions than one can answer in an individual career. In other words, get the right question, but don't spend so much time figuring out what the right question is that you don't move forward. The questions should lead you to research that have the potential for being fun. While securing NIH funding is an important milestone in your career, remember if your study is funded, you will be doing it for at least the next 2–5 years and it will impact your future area of research. Don't propose any research question that you really do not think you will enjoy for the "long term". Aside from the fun aspect (which is an important one), the "right" research question should lead to a hypothesis that is testable, that is based upon existing knowledge and fills and existing gap in specific areas of knowledge. Finally, the "right" research question is a question

Table 19.2 Components of an abstract

The research question that the study will address
A brief justification to orient the reviewer
The overall hypotheses to be tested
The study population to be recruited
The methods you will use
The overall research plan
How the proposed research, if, successful, will advance your field of research

that can be transformed into a feasible study plan. How does one find the "right" research question? Open your eyes and observe: patients often provide clues into what is known and unknown about clinical practice. This approach formed the basis of one of the authors R01 ("does the variable left ventricular hypertrophy response in the context of hypertension have a genetic basis?"). Another way of coming by the "right" research question is through teaching and through new technologies.

Abstract

The abstract and specific aims (described below) are the two most important components of any grant application and must provide a cohesive framework for the application. The abstract provides an outline of the proposed research for you and the reviewer. Include in the abstract the research question that the study will address with a brief justification to orient the reviewer, the overall hypotheses to be tested, the study population you will recruit, the methods you will use, and the overall research plan (Table 19.2). These details are important so that study section personnel can decide which study section best fits the grant. The final statement in the abstract should indicate how the proposed research, if, successful, will advance your field of research. Always revise the abstract after your complete proposal has been written so that it agrees with what you have written in the research section.

Developing a Research Question and Specific Aims

In developing a research question, one needs to choose a "good" or the "right" question as discussed above (also see Chap. 2). The "right" research question should lead you towards a testable hypothesis about the mechanisms underlying the disease process you are studying. A testable hypothesis will also require a feasible experimental design such that you can test the various predictions of your hypotheses in the most rigorous way so that your study does all that it can to fail to refute the null hypothesis if it is true. Once you have a testable hypothesis and a feasible and

Table 19.3 Components of specific aims

Components	Example
A brief introduction that underscores the importance of the proposed research	LVH is a common condition associated with cardiovascular morbidity and mortality…
The most important findings to date	We have shown that LVH is, at least in part, genetically determined…
The problem that the proposed research will address	We anticipate these strategies will identify genetic variants that play clinically significant roles in LVH

rigorous design to translate the research question into the hypothesis, there are certain necessary components that one needs to consider. Certainly, the hypothesis should define the study purpose, but should also address: the patient/subject eligibility (i.e., characterize the study population); the exposure (or the intervention); the comparison group; and the endpoints (outcomes, dependent variable – refer to PI(E) COS in Chap. 3). As described by Hulley et al. the criteria of a good hypothesis is that it is feasible, interesting, novel, ethical, manageable in scope, and relevant. It is helpful to engage colleagues to respond to how novel and interesting the hypothesis is and to address whether the results of your study will confirm extend, or refute prior findings, or provide new knowledge. Arguably, the most common mistake a new investigator makes is to have failed to narrowly focus the question such that it is feasible to answer with the research proposed. That is, avoid having a question that is too broad or vague to be reasonably answered. Finally, include only experiments that you and your colleagues and you're your institution have the expertise and resources to conduct.

For the NIH grant, the hypotheses are written in Section A of the proposal, named "Specific Aims." Specific aims are extensions of your research questions and hypotheses, and they should generally be no more than one page and should include (i) a brief introduction that underscores the importance of the proposed research, (ii) the most important findings to date, and (iii) the problem that the proposed research will address. Using the example of the genetic determinants of ventricular hypertrophy mentioned above, the aims section began with "(i) LVH is a common condition associated with cardiovascular morbidity and mortality… (ii) we have shown that LVH is, at least in part, genetically determined…. (iii) we anticipate these strategies will identify genetic variants that play clinically significant roles in LVH (Table 19.3)". Such knowledge may suggest novel pathways to be explored as targets for preventive or therapeutic interventions.

Even though the specific aims should be comprehensive in terms of the proposed research, the aims should be brief, simple, focused, and limited in number. Draft the specific aims like you would a novel such that you create a story that builds logically (i.e. each aim should flow logically into the next aim). The aims should be "realistic", that is, they should represent one's capacity for completing the work you propose and within the budget and the time requested. Use a variety of action verbs, such as characterize, create, determine, establish, delineate, analyze, or identify, to

Table 19.4 What should be in the background and significance section

What is the current state of knowledge
Why is this research question important
What gaps in knowledge will this project fill
Does it fill a specific gap in knowledge
More generally, why is this line of research important

name a few. Most importantly, keep the aims simple, at the appropriate level of your team's expertise, and where you have supporting preliminary data.

Writing specific aims can take on a variety of models. One model might be to have each aim present a different approach that tests a central hypothesis. Another model may be to have each aim develop or define the next logical step in a disease process. You should avoid a model in which an aim is dependent of the successful completion of an earlier aim. In other words, do not have aims that could only successfully move when and if the earlier aim is successful. Such contingent aims reduce the scientific merit of the grant since reviewers cannot assess their probability of success.

The Background and Significance Section

The background and significance section must convince your reviewers that your research is important; in other words, you must market your idea to reviewers in such a way that it engages them intellectually and excites them in terms of the potential for impact on clinical practice, and ultimately, health. You must also provide the foundation for your research, and show your knowledge of the literature. To provide the reviewer evidence of your ability to critically evaluate existing knowledge, the background and significance section should not only clearly state and justify the hypotheses, but should also justify variables and measurements to be collected, and how the research will extend knowledge when the hypotheses are tested. The wrap-up paragraph should discuss how your proposed research fits into the larger picture and demonstrate how the work proposed fills an important gap in knowledge. Some key questions to address are (Table 19.4):

- What is the current state of knowledge in this field?
- Why is this research important? Does it fill a specific gap in knowledge?
- What gaps in knowledge will this project fill?
- More generally, why is this line of research important?

Captivate the reviewer by emphasizing why the research question is fascinating. For instance, what is known? What question is still unanswered? And why do we want to answer this particular question? Finally, you must address what your proposed project has to do with the public health or clinical medicine.

Background and significance sections will be read by experts in your field since reviewers are selected based on their matched expertise with your project. Therefore, you must be both factual and provide "readable" material. Whenever possible, use cartoons or diagrams to clarify concepts and to visually break up the page. It is also useful to create a "road map" for your application in the introductory paragraph (e.g. in one of the author's section, the following was used: "in this section, we review (1) the epidemiology of hypertension; (2) the pathophysiology of hypertension; (3) other medical consequences of hypertension; (4) the clinical treatment of hypertension; (5) the genetics of hypertension, and (6) implications for proposed research". Having this roadmap is particularly important for the reviewer, since often a busy reviewer may only skim headings. Your headings within the background and significance section should lead the reviewer to know fully why that section is in the application. Like the specific aims, it is important to keep the background and significance section simple, to avoid jargon, to define acronyms, to use "sound bites", and repeatedly use these "sound bites" throughout the application. Finally, engage a colleague from a close but unrelated field to read the background section to test the ease of understanding of its structure and content to a non-expert.

Preliminary Studies Section

The best predictor of what you will do tomorrow is what you did yesterday

The NIH has specific Instructions for the preliminary studies section, and "suggest" this section should provide an account of the principal investigator's preliminary studies relevant to the work proposed and/or any other information—from the investigator and/or the research team—that will help to establish their experience and competence to pursue the proposed project. Six to eight pages are recommended for this section. Content should include previous research, prior experiments that set the stage for the proposal and build the foundation for the proposed study. The pilot data provided should be summarized using tables and figures. Interpretation is also important so that you demonstrate your ability to articulate accurately the relevance of your pilot data and correctly state the impact of your prior work. In a related way, this section also uses the previous results to demonstrate the feasibility of your proposed project. To convince reviewers of your research feasibility, you should discuss your own work--and that of your collaborators - on reasonably related projects, in order to convince reviewers that you can achieve your research aims. Pilot studies are required for many (but not all) R-series grants, and are extremely important to show your project is "do-able".

The preliminary study section is particularly important for junior investigators where there may be inadequate investigator experience or training for the proposed research, a limited publication record, and/or a team that lacks the skill set required for the research proposed. The quality of the preliminary study section is critically important for junior investigators as the quality of the presentation of the pilot work is evidence of your ability to complete the work you propose.

Research Design and Methods

The research design and methods section is the place where you cover all the materials and methods needed to complete the proposed research. You must leave adequate time and sufficient space to complete this section. Many applicants run out of time and page requirements before the last aim is addressed in sufficient detail, significantly weakening the application. As concordant with the aims, it is important to not be overly ambitious. In the opening paragraph of this section it is also an important time to re-set "the scene" by refreshing the reviewer regarding the overview for each specific aim. Sometimes, this is the section where reviewers began to read the application. As you progress, use one paragraph to overview each specific aim, and then to deal with each sub-aim separately.

You should be clear, concise, yet detailed regarding how you will collect, analyze, and interpret your data. As stated in the specific aims section, it is important to keep your words and sentence structure simple because if the reviewer is confused and has to read your proposal numerous times, your score will suffer. At the end of this section give your projected sequence or timetable. This is the section to convince reviewers that have the skills, knowledge and resources to carry out the work, and that you have considered potential problems and pitfalls and considered a course of action if your planned methods fail. Finally, by providing data interpretation and conclusions based on the expected outcome, or on the chance that you find different results than expected (a not uncommon occurrence), it demonstrates that you are a thoughtful scientist.

One should provide a bit of detail for each section, such as addressing the design chosen for your research project and why you chose that design rather than another, what population you will study and why, what will be measured and how it will be operationalized in the clinical setting, and on what schedule. Develop each specific aim as a numerical entity by reiterating it, and using **BOLDING** or a text box in order to highlight it. Briefly re-state the rationale for your each aim.

Patient Enrollment

Convey to the reviewer your appreciation for the challenges in recruiting. Discuss from where the population will be recruited, what the population characteristics (gender, age, inclusion and exclusion criteria) will be, how subjects will be selected and the specific plans for contact and collaboration with clinicians that may assist you. Provide any previous experience you have with recruitment and include some numbers of subjects, and response rates, from previous or preliminary studies. Provide strategies to remedy any slow recruitment that might occur. Be cognizant of

NIH policies in order to properly address issues related to gender, minority, and children inclusions and exclusions.

One also needs to consider and address the participant burden for the proposed research in order to properly weigh the benefits and costs of participation. In many studies, research subjects should be paid but not to the degree that it is coercive (See Chap. 8).

Methods

One should provide details for the most important techniques to be used in your research. For commercially available methods you need only to briefly describe or reference the technique; but, for methods crucial to your aims, you need to provide adequate description such as referencing published work, abstracts, or preliminary studies.

In the author's experience, there are some common weaknesses of the Methods Section. These weaknesses include such issues as an illogical sequence of study aims and experiments; that subsequent aims (also known as contingent aims) rely on previous aims such that if the previous aims fail, the study comes to a halt. Inadequate description of contingency plans, or poorly conceived plans, or plans that are not feasible significantly weaken a proposal. Other weaknesses include not adequately describing or constructing the control groups; and/or underestimating the difficulty of the proposed research.

Tips for Successful Grants

A successful grant proposal generally "tells a story" and engages the reviewer. The proponent should anticipate questions that are likely to occur and present a balanced view for the reviewers. To be successful, you must not take things for granted, and you must deliver a clear, concise, and simply stated set of aims, background, preliminary studies, and experimental methods that has addressed threats to both internal and external validity. You must be able to follow directions precisely and accurately, and target your grant to the expected audience (i.e., your reviewer). Your timeline and budget must align with your aims. As stated earlier, you should obtain an independent review both from your mentors and collaborators, but from external reviewers if possible. And finally, and perhaps most importantly, remember, not every proposal gets FUNDED!, in fact only a minority get funded so it is prudent to submit a number of different proposals, understanding that you won't get funded unless you submit proposals. When resubmitting proposals you should be careful to revise it based upon the critique and realize that reviewers are attempting to help you make your study better. There is no use getting mad–get funded instead! Every application must be above any level of embarrassment (i.e., do not submit anything that is not your best work). Develop a game face after submission, and be confident about your proposal. To maintain your sanity through the process, convince yourself that your grant won't get funded while concurrently reminding your colleagues it is tough to get funded.

Types of NIH Research Funding

There are a number of types of NIH research funding, but of most relevance to clinical research are:

Grant (Investigator Initiated)
Cooperative Agreement (NIH is a partner; assistance with substantial involvement)
Contract (purchaser)
Training Awards
Research career development awards
Mentored NIH Career Development Awards
K01/K08 Research/Clinical Scientist
K23 and K24 Patient Oriented Research
Mentored Research Scientist Development Award (K01)

These awards provide support for an intensive, supervised career development experience, leading to research independence for early or mid-career training, as well as to provide for a mechanism for career change (K24). The K24 requires that the applicant have a substantial redirection, appropriate to the candidate's current background and experience, or that the award provides for a significant career enhancement. "Unlike a postdoctoral fellowship, the investigator must have demonstrated the capacity for productive work following the doctorate, and the institution sponsoring the investigator must treat the individual as a faculty member."

The characteristics of the ideal candidate may vary. For example, the candidate may have been a past PI on an NIH research or career development award; but, if the proposed research is in a fundamentally new field of study or there has been a significant hiatus because of family or other personal obligations, they may still be a candidate for one of these awards. However, the candidate may not have a pending grant nor may they concurrently apply for any other career development award.

Summary

Remember; logically develop your aims, background, preliminary studies and research design and methods into a cohesive whole. Clearly delineate what will be studied, why it is important, how you will study it, who(m) you will study, and what the timeline is to complete the research. When writing, say what you're going to say, then say it, and finally summarize what you said. Write a powerful introduction, particularly if you are constructing a revised application. Develop your "take-home messages" and reiterate them throughout your application. Finally, be tenacious: learn from your mistakes, pay careful attention to critiques, collaborate with smart people and find a good mentor. And, above all, keep it simple.

Manuscript Preparation

Many manuscripts follow after work is presented in abstract form at a major medical meeting. But, Fosbol has noted “while conferences allow abstracts public airing and media attention, we find it perplexing that two-thirds of these abstracts will not be published within a 2-year period” [2], and only 40 % will be published at 5 years. Fosbol also pointed out that abstracts rejected for presentation still had a 1 in 4 chance of being published; and, Winnik et al. found that among abstracts accepted to the European Society of Cardiology the subsequent publication of a manuscript reached 38 % and for rejected abstracts 24 % [3]. To analyze this issue further, Krzyzanowska et al. [4] reported on identifying factors associated with time to publication. They found that of 510 randomized trials, 26 % were not published in full within 5 years after presentation at a meeting. Eighty-one percent of the studies with significant results had been published but only 68 % with non-significant results were published in this same time period. They stated “non-publication breaks the contract that investigators make with trial participants, funding agencies, and ethics boards”.

The quality of reporting of abstracts is another issue that has been examined. Krzyzanowska et al. evaluated 510 abstracts and reported deficiencies in almost all [5]. For example 22 % of the abstracts failed to provide explicit identification of the primary endpoint. The general recommendations for abstract content are shown in Table 19.5.

There are many areas of overlap between writing a grant and writing a manuscript, but many differences as well. Irrespective of whether one is writing a grant or a manuscript (or anything else for that matter) it is important to remember that your writing is a reflection of your thinking, and as such, it should be clear and concise. If you want to be taken seriously, one must become a better writer (and that applies to all of us). Kerpan [6] outlined five steps to become a better writer as follows: practice, practice, practice; say it out loud; make it more concise; work on your headlines, and read great works.

Table 19.5 Suggested guidelines for what should be included in an abstract

Abstract guidelines
Reported if space permits
Dates of accrual
Description of statistical analysis
Whether ITT was used if an RCT
Patient attrition
Pre-specified secondary and/or subgroup analyses
Should not be reported
Results of secondary analyses not pre-specified
Results of subgroup analyses not pre-specified

Table 19.6 Pneumonic for helping to remember a structured approach for framing questions

PICOS or PECOS
The patient population or disease being addressed (P)
The interventions (I) or exposure (E)
The comparator group (C)
The comparator group (C)
The study design chosen (S)

The basic outline of a manuscript and a grant are the same: Title, Abstract, Introduction, Methods (to include patient recruitment, characteristics of the study population, study design etc.), and Statistical Analysis. Unlike a grant, however, save for preliminary study results, the actual results of the study are then presented followed by a focused Discussion, which should include study limitations, and finally, conclusions. Obviously budgetary data, data sharing considerations, and a few other issues peculiar to grants are not part of the manuscript preparation; but, funding sources and potential conflicts of interest (see below) should be listed.

In general, formulating relevant and precise questions that can be answered can be complex and time consuming. A structured approach for framing questions that uses five components may help facilitate the process. This approach is commonly known by the acronym "PICOS or PECOS" (Table 19.6): the patient population or the disease being addressed (P), the interventions (I) or exposure (E), the comparator group (C), the outcome or endpoint (O), and, the study design chosen (S). Providing information about the population requires a precise definition of a group of participants (often patients), such as men over the age of 65 years, their defining characteristics of interest (often disease), and possibly the setting of care considered, such as an acute care hospital. The interventions (exposures) under consideration in the manuscript need to be transparently reported. For example, if the reviewers answer a question regarding the association between a woman's prenatal exposure to folic acid and subsequent offspring's neural tube defects, reporting the dose, frequency, and duration of folic acid used in different studies is likely to be important for readers to interpret the review's results and conclusions. Other interventions (exposures) might include diagnostic, preventative, or therapeutic treatments, arrangements of specific processes of care, lifestyle changes, psychosocial or educational interventions, or risk factors. Clearly reporting the choice of the comparator (control) group, and the intervention(s), such as usual care, drug, or placebo, is essential for readers to fully understand the reasons for one's choice. Comparator groups are often very poorly described as are the reason(s) for that choice. Clearly reporting what the intervention or exposure is compared with is very important and may sometimes have implications. The outcomes of the intervention being assessed, such as mortality, morbidity, symptoms, or quality of life improvements, should be clearly specified as they are required to interpret the validity and generalizability of the studies results.

Guideline Statements for Manuscript Preparation (Table 19.7)

Guideline statements for manuscripts include:

CONSORT (Consolidated Standards of Reporting Trials)
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology)
PRISMA (Preferred Reporting items for Systematic Reviews and Meta-Analyses) that is an extension of QUOREM (Quality of Reporting of Meta-Analyses for meta-analyses of RCTs)
MOOSE (Meta-analysis of Observational Studies in Epidemiology, for meta-analyses of observational trials)
STREGA (Strengthening the Reporting of Genetic Association Studies - an extension of STROBE)

A checklist has been formulated for each, and these guidelines have been accepted by many (most) Journals. These checklists vary somewhat from each other but there are many areas in common as well.

CONSORT [7]

The CONSORT 2010 Statement is a 25-item checklist and a flow (exclusionary cascade) diagram (see Fig. 19.1). It provides guidance for reporting all randomized controlled trials, but focuses on the most common design type—individually randomized, two-group, parallel trials. Other trial designs, such as cluster randomized trials and non-inferiority trials, require varying amounts of additional information.

Title

Often, the title of the manuscript is added just before the manuscript is submitted to a journal for their consideration for publication, and yet it is the first thing that the editor and reviewers will see. Therefore, the title for the manuscript should be given some thought. A catchy title might grab the interest of the potential reader, but it

Table 19.7 Examples of some guideline statements for what should be included in manuscripts

CONSORT	Consolidated Standards of Reporting Trials
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
PRISMA	Preferred Reporting items for Systematic Reviews and Meta-Analyses
MOOSE	Meta-analysis of Observational Studies in Epidemiology
STREGA	Strengthening the Reporting of Genetic Association Studies

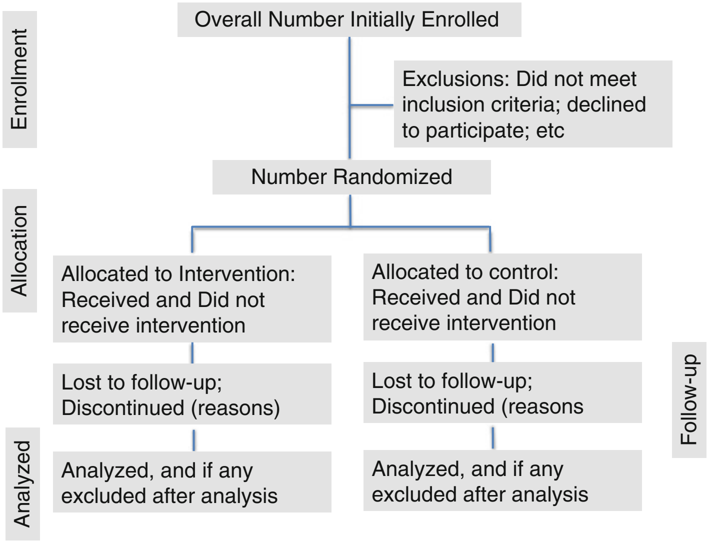

Overall Number Initially Enrolled
Enrollment
Exclusions: Did not meet inclusion criteria; declined to participate; etc
Number Randomized
Allocation
Allocated to Intervention: Received and Did not receive intervention
Allocated to control: Received and Did not receive intervention
Lost to follow-up; Discontinued (reasons)
Lost to follow-up; Discontinued (reasons
Follow-up
Analyzed
Analyzed, and if any excluded after analysis
Analyzed, and if any excluded after analysis

Fig. 19.1 Flow diagram of the progress through the phases of a parallel randomized trial of two groups (that is, enrolment, intervention allocation, follow-up, and data analysis)

should be accurate and reflect what was actually addressed by the study. It is often surprising how frequently the two are disparate. An example of this disparity (a true example with some words changed to protect the author) is a manuscript entitled "Discrepancy of Drug X and Drug Y Between the Blood Pressure Lowering Effect and Effect on Endothelial Function", and the studies conclusion which stated "in conclusion, our results suggest that Drug X is recommended as second-line treatment despite the failure to lower blood pressure as much as Drug Y." Finally, whereas the title of an oral presentation might include alliterations, these should be avoided in manuscript submissions (in the authors opinion Editors have little sense of humor).

Abstract

After the title, the Abstract is the next thing editors, reviewers, and ultimately readers will see. In fact, sometimes it is the only thing about the manuscript that will be read. It frequently is also what will be electronically accessible. Thus, like the title, considerable thought should be given to its content. Most journals are now suggesting and even requiring a structured abstract. This begins with the Background of the Study or Study Objectives depending on the specific journal. The Methods Section is usually next followed by a Results Section, and finally Study Conclusions. For most journals the word count for Abstracts ranges from 200 to 350 so one needs to carefully read the Instructions to Authors Section for details. Many (most) journals are now using an online submission format, so if the word count is in excess of what the journal allows it will automatically prevent submission. Thus, the abstract needs to be on the one hand concise, but on the other hand include all pertinent aspects of the design and results. Frequently it takes longer to write the abstract than the manuscript which brings to mind a quote by the French philosopher and mathematician Blaise Pascal who is quoted as saying *"I am sorry to have wearied you with so long a letter but I did not have time to write you a short one"* [8].

Keywords

Often, not much thought is given to keyword selection, and yet it is these words that will allow for future searches to identify the appropriate studies for literature searches and meta-analyses. One should, in fact, give due thought to these words by considering what you would want to enter into your search engine to find the data included in one's manuscript.

Introduction

The manuscript introduction should be compared to the grant's background and significance section but briefer. It should set the stage for the aims and/or hypothesis for the study one is writing about, thus, again it should be relatively brief and focused, that is, it should not be a literature review. It should also state the aim, hypothesis, and/or objective of the study about which one is reporting. The main function of the introduction though, is to "motivate the audience to read the paper and care about its results" [9].

Methods

The Methods section includes discussions of the trial design (e.g. parallel, factorial, crossover etc.) including the allocation ratio, eligibility criteria, the settings and location of the study population, and intervention details (e.g. how and when administered) with enough detail to allow replication. In addition, outcome(s) should be completely defined, pre-specified, and include both primary and secondary outcomes and how and when they were assessed. A statistical section should include sample size calculations and any planned interim analyses and stopping guidelines, randomization methods, type, etc., blinding (method and who was blinded), and statistical methods. In addition, the analytical approach used (e.g. intention to treat, etc.), should be included. If subgroup analyses are performed, the numbers of such analyses and whether they were pre-specified or *post hoc* should be mentioned. In addition, a discussion should be included of how patients/subjects who are lost to follow-up were handled.

Results

The results section should include a participant flow diagram (so that the reader can assess the studies generalizability, and other potential biases, recruitment strategies, baseline data (i.e. a description of the baseline features important to the study), almost always this includes demographics (e.g. comparing age, sex, socioeconomic differences between the groups studied). This baseline table is frequently called "table 1" of most manuscripts. The collection of baseline data has at least four main purposes:

- To characterize the patients included in the trial, i.e. to determine how successful randomization was
- To allow assessment of how well the different treatment groups are balanced,
- To allow for analysis per treatment group,
- To allow for subgroup analysis in order to assess whether treatment differences depend on certain patient characteristics

Some questions raised by baseline data analysis are; how is it measured? What does it mean if there is or is not statistically significant differences? And, does sample size matter? [10] An argument that exists is over whether to use statistical testing of baseline differences or to rely on a subjective comparison of baseline variables. One side of the argument is that on the one-hand, just because there is a difference in a baseline variable it doesn't mean that it influences the outcome(s); and, on the other hand, just because there is no statistical difference doesn't mean that there is not a baseline variable that does influence outcome. Furthermore, if the sample size is large, small differences that may not be clinically meaningful might show very significant statistical differences. Irrespective, it is generally agreed that statistically significant differences or lack thereof should not be completely relied upon, statistically significant differences are less of an issue for sample sizes over 500, and that baseline variables give some measure to assess comparability between the groups under study. Table 19.8 is an example of the "table 1" baseline comparability's. This table can also be used to illustrate the issue of "column vs. row' percentages and how data is displayed. If one is interested in emphasizing how a variable is distributed over outcomes (i.e. the percentages of each outcome per group) the data would be portrayed one way. On the other hand, if the interest is in emphasizing the percentages of groups that have the outcome, the data would be portrayed in another way (See Table 19.9).

In addition, the numbers analyzed for each group, estimated effect size and its precision (both absolute and relative effect size) any ancillary analyses, and any safety or analyses for harm.

Discussion

The discussion section should begin with a summary of results presented in general terms. Next should be a focused discussion of the study results in terms of

Table 19.8 An example of a baseline variables table

Table 1. Pre-hypertension analysis cohort, REGARDS, N=24,393

Baseline characteristics by different classes of hypertension,

			Pre-Hypertension (n=6,066)		Hypertension (n=13,742}		
	All participants (N=24, 39 3)	Normotensive (n=4,585), (18.8 %)	Pre-HTN1 (n=4,000) (16.4 %)	Pre-HTN2 (n=2,066) (8.5 %)	Not Controlled (n=5,364) (22.0 %)	Controlled (n=8,378), (34.4 %)	P value
Demographics							
Age, years, *M*(*SD*)	64.1 (9.3)	61.0 (9.1)	62.8 (9.3)	64.6 (9.3)	65.9 (9.3)	65.2 (9.0)	<.001
Gender, (%)							<.001
Male	41.6	37.0	47.7	50.3	45.5	36.5	
Female	58.4	63.0	52.3	49.7	54.5	63.5	
Race, (%)							<.001
Black	42.4	24.9	31.6	37.3	54.3	50.7	
White	57.6	75.1	68.4	62.7	45.7	49.3	
Region, (%)							<.001
Belt	34.7	33.8	32.9	34.6	36.1	35.2	
Buckle	20.9	21.9	20.8	17.8	18.9	22.5	
Nonbelt	44.4	44.3	46.3	47.7	44.9	42.4	
Education, (%)							<.001
Less than high school	11.6	6.2	8.1	11.3	16.2	13.3	
High school only	25.4	21.6	23.1	25.3	28.1	26.9	
Some college or College graduate	63.1	72.3	68.8	63.5	55.8	59.8	
Annual income, (%)							<.001
$20K or less	19.6	12.3	15.1	17.3	25.8	22.0	
All other	80.4	37.2	85.0	82.7	74.2	78.0	

what is already in the literature, to include similarities and differences. If mechanisms have been explored or suggested by the study, a discussion should include that as well.

Every study has limitations, so a frank discussion of those limitations, and the degree to which they might alter the results is appropriate. This should include any sources of potential bias and the generalizability of the study results.

Table 19.9 Compares the presentation of column vs. row data

Column vs. Row % comparing BP classes					
	No HTN	preHTN	Controlled HTN	Uncontrolled HTN	Total n
Black	26.2	36.2	50.7	54.3	10331
White	73.8	63.8	49.3	45.7	14057
Male	40.0	49.8	36.5	45.5	14251
Female	60.0	50.2	64.5	54.5	10137
Total n (%)	6791 (100 %)	3860 (100 %)	8378 (100 %)	5359 (100 %)	
	No HTN	preHTN	Controlled HTN	Uncontrolled HTN	Total n (%)
Black	17.2	13.5	41.1	28.2	10331 (100 %)
White	35.7	17.5	29.4	17.4	14057 (100 %)
Male	26.8	19.0	30.2	24.1	14251 (100 %)
Female	28.6	13.6	37.3	20.5	10137 (100 %)
Total n	6791	3860	8378	5359	

The top table addresses the % of subjects with No Hypertension (HTN), prehypertension (preHTN) etc who are Black, White, Female, and Male; while the bottom table addresses what % of Blacks have No HTN, preHTN etc

Table 19.10 An outline of how to construct the discussion section of a manuscript

Paragraph #	What the paragraph should include
1	Describe the major findings and answer the research question
2	Interpret and explain the major findings
3–5	Compare the results with the literature and highlight literature that conflicts with the findings
6	Discuss the study limitations and its generalizability
7	Discuss unanswered questions and propose further research
8	Make conclusions supported by the findings and consistent with the manuscripts title

Finally the manuscript should end with a focused conclusion that is again reflective of the study title and aims, followed by any acknowledgements, potential conflicts of interest, and funding sources. Welch [9, 11] have published outlines of what should go into the discussion and in what order and this is reproduced in Table 19.10.

In a humorous but appropriate list of rules developed by Frank l. Vasco entitled "How to Write Good" [12], there are 23 tips provided as follows: avoid alliterations **always**; prepositions are not words to end a sentence with; avoid clichés like the plague (they are old hat); employ the vernacular; eschew ampersands & abbreviations etc.; parenthetical remarks (however relevant) are unnecessary; it is wrong to ever split an infinitive; contractions aren't necessary; foreign words and phrases are not *apropos*; one should never generalize; eliminate quotations (as Ralph Waldo Emerson once said, "I hate quotations. Tell me what you know"); comparisons are as bad as clichés; don't be redundant, don't use more words than necessary, it's highly superfluous; profanity sucks; be more or less specific; understatement is always best; exaggeration is a billion times worse than understatement; one word sentences…: eliminate analogies in writing, they are like feathers on a snake; the passive voice is to be avoided; go around the barn at noon and avoid colloquialisms; even if a mixed metaphor sings, it should be derailed; who reads rhetorical questions. Added to this is the proper choice of words to realistically reflect what you as the author is really saying. *Accad* [13] makes this point by suggesting that authors have embraced the activity of fortunetelling with the increasing use of the word "predicts" in medical writing which he sites a hyperbole. His point is that the use of the much mis-understood P value (to provide a sense of objectivity) refers to a group effect and not an individual patient. As an example of this latter concept he points out that when one "…is told that cardiac troponin predicts death because its elevation in the postsurgical setting is more prevalent among those who later died" when the actual results were when elevated values were identified 21 % died vs. 6 % who lived (and this ignores the fact that in this particular group, elevated levels could foretell a fatal outcome in only 32 %).

STROBE [14]

The STROBE statement defines the scope of the recommendations that cover three main study designs: cohort, case-control, and cross-sectional studies. A checklist of 22 items has been developed that relate like the CONSORT statement to the title, abstract, introduction, methods, results, and discussion Section. 18 items are common to all three observational study designs, and four are specific for cohort, case-control, or cross-sectional studies.

Otherwise, the same or similar principles hold for STROBE and CONSORT. Some differences between STROBE and CONSORT relate to the study designs. For example, for cohort studies the matching criteria and number of exposed and unexposed subjects should be mentioned, while for a case-control study the matching criteria and the number of controls per case should be emphasized. In terms of statistical analyses those used for control of confounding should be described as well as how missing data was addressed (see Chap. 3) along with a description of any sensitivity analyses.

PRISMA [15]

The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of RCTs again has many common features of the other guideline statements and consists of a 27-item checklist and a four-phase flow diagram. Some differences between the guidelines include the mention in the title that identifies the report as a systematic review, meta-analysis, or both; and a mention of the synthesis and search methods in both the abstract and methods sections. In the methods section one should indicate if a review protocol exists, if and where it can be accessed (e.g., web address), and, if available, provide registration information including registration

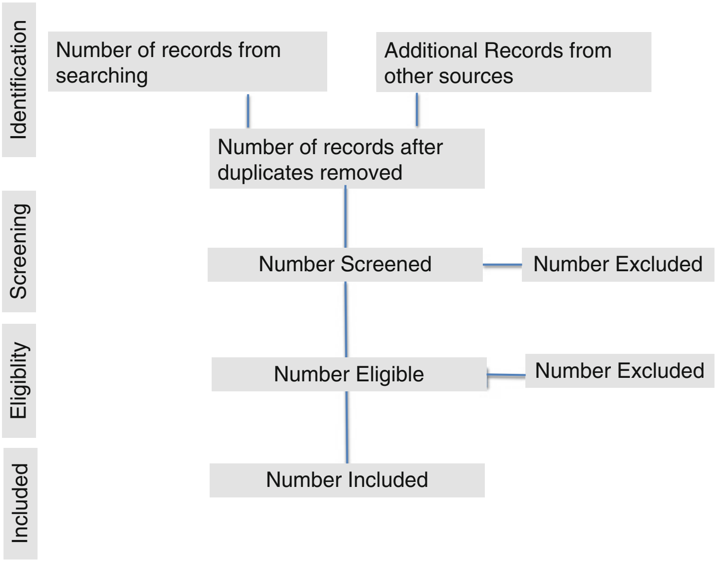


Fig. 19.2 The flow of identifying and choosing studies to be included in the meta-analysis

numbers. Eligibility in the context of meta-analyses specifies the characteristics of the studies included in the analysis (e.g., length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale. Information sources should be described (e.g., databases with dates of coverage, contact with study authors to identify additional studies) the date last searched, and one should present a full electronic search strategy for at least one database, including any limits used, such that it could be repeated. State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis). More specifically, the method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators should be mentioned. The risk of bias in the individual studies that make up the meta-analysis and the methods used for assessing those risks (including specification of whether this was done at the study beginning, or outcome level), and how this information was to be used in any data synthesis. One should present the numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram (Fig. 19.2). The results section should include for each study, the characteristics for which data were extracted (e.g., study size, follow-up period) and the results of individual studies. For all outcomes considered (benefits or harms), there should be a presentation for each study that includes: (a) simple summary data for each intervention group and (b) effect estimates and confidence intervals, ideally with a forest plot, including confidence intervals and measures of consistency.

For meta-analyses of observational trials the Meta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines are suggested [16].

STREGA [17]

The STREGA statement for reporting genetic association studies builds on the STROBE guidelines and provides additions to 12 of the 22 items on the STROBE checklist. The components of the title, abstract, study design and population are similar to STROBE. But, things unique to genetic studies (e.g. whether the Hardy-Weinberg equilibrium was considered, methods used for genotypes or haplotypes, reporting of the numbers of individuals in whom genotyping was attempted and the numbers in whom it was successful) are obviously necessary for this specialized field. The interested reader can refer to the guideline document for more details and Chap. 11.

Conflicts of Interest, Authorship, Coercive Citation, and Disclosures in Industry-Related Associations

The International Committee of Medical Journal Editors (ICMJE) have published authorship criteria and to summarize "authorship credit should be based upon (1) substantial contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (2) drafting the article or revising it critically for important intellectual content; and (3) final approval of the version to be published". Authors should meet all three criteria. Regarding authorship, there has been a war on so called "ghost authorship". According to one report, in 2008, honorary authors were attached to 25 % of research reports, 15 % of review articles, and 11 % or Editorials published in six major journals; [18] while in another report Mowatt et al. 39 % had evidence of honorary authorship [19] Greenland and Fontanarosa note that many times honorary authorship amounts to "coercive authorship" in which a senior person insists on being listed as an author even though they did not contribute substantially to the work; while in other cases the senior author is added in the hopes of increasing the chance of a manuscript being accepted [20]. Ghostwriting is defined as a person who writes books, articles, stories, reports, or other texts that are officially credited to another person-the opposite of honorary authorship. However, in the medical arena, what really happens is a ghostwriter submits their work to a medical investigator who then has the chance to edit, delete or add to the text as they see fit. None-the-less ghostwriting is highly discouraged in the scientific literature. Honorary and ghost authorship are frequently lumped together as "inappropriate authorship".

Coercive citation is the practice in which an editor "forces" an author to add citations to an article (usually from that Editor's Journal) before they will agree to publish it. This is done to inflate the journal's impact factor (IF). Wilhite and Fong noted that despite the IFs shortcomings they continue to be a means by which the quality of science is weighed [21]. The Impact Factor of a journal was devised as a way to rank scientific journals, and is a measure of how often, on average, papers published in a journal are cited in other academic publications. IFs are used in some institutions as a promotional tool, but more recently IFs have become a source of increasing controversy, and Franck has criticized the practice where "success in science is rewarded with attention" [22].

The Institute of Medicine defines conflict of interest (COI) as "*a set of circumstances that creates a risk that professional judgment or actions regarding a primary interest will be unduly influenced by a secondary interest*". The term COI has taken on an almost presumption of guilt, partially the result of a few highly publicized incidents in which there was an attempt to manipulate clinical research by blocking publications, withholding data, and falsely reporting results of a 12-month study as a 6 month trial. These events led in 2004 to the ICMJE's call for mandatory clinical trial registration [23] –(this reference also serves as an excellent in-depth review of the subject).

There is a good deal of variation between journals in what information they require before accepting manuscripts for publication. One journal requires a 17-page questionnaire to be filled out. This has resulted in attempts to develop a uniform disclosure form, but with little success. In this regard, the authorship issue involved with industry-supported studies highlights the conflicts between academia and Industry. The general view is such funded studies particularly those with industry authors would be more biased and of lesser quality that studies funded through other sources. The increasing number of clinical trials that have full or partial industry funding has been increasing, and industry employees are increasingly appearing as coauthors of clinical trials that adds fuel to this belief, and yet there is little proof to support that belief. Booth et al. evaluated reports of RCTs evaluating systemic therapy of breast, colorectal and non-small cell lung cancer [24] and found that for-profit sponsorship and statistically significant results are independently associated with the endorsement of the experimental arm, even though authors who perform key roles in the conception, design, analysis, and interpretation of oncology trials are likely to have financial ties to industry [25]. Kaiser et al. published a non-industry supported study entitled "Is Funding Source Related to Study Design Quality in Obesity or Nutrition Supplement Randomized Control Trials (RCTs)?" The purpose of that study was to examine systematic quality differences amongst obesity and nutrition RCTs based on funding status in top tier journals. Thirty-eight obesity or nutrition intervention RCT articles were selected from high-profile journals (Lancet, Annals of Internal Medicine, JAMA, British Medical Journal) published between 2000 and 2007. Paired papers were selected from the same journal published in the same year, one not reporting industry funding and the other reporting industry funding. Papers had the following identifying information redacted: journal, title, authors, funding source and institution(s). Three raters independently and blindly rated each paper according to the Chalmers Method [26]. Total quality scores were calculated. The Wilcoxon signed ranks test and paired-samples t-test were used to compare Chalmer's Index score between industry-funded versus non-industry funded studies. Inter-rater reliability using an intraclass correlation coefficient = 0.82 (95 % C.I. = .80 – .84). Mean quality score for industry-funded studies = 13.7, SD = 3.01; for non-industry funded studies mean score = 13.2, SD = 4.09. The

Table 19.11 Descriptive and test statistics for total and subscale scores for each funding category

	Industry funded studies (M, SD, n = 19)	Non-industry funded studies (M, SD, n = 19)	Wilcoxon signed ranks test (two-tailed)
Overall total quality score	84.5, 7.04	79.4, 13.00	p = .334
Study protocol score	50.4, 6.25	46.3, 11.13	p = .331
Statistical analysis score	25.2, 2.68	24.5, 2.87	p = .450
Presentation of results score	8.9, 2.03	8.6, 2.18	p = .553

From: Kaiser et al. [27]

Wilcoxon Signed Ranks test statistic, $Z = -.523$, $p = .601$ (two-tailed) indicated no categorical difference in study quality. Paired-samples t-test also indicated no significant mean difference in total quality scores between funding categories, $t(18) = .587$, $p = .564$ (two-tailed). They concluded that recently published RCTs in nutrition and obesity that appear in top-tier journals seem to be equivalent in quality of design, regardless of funding source (Table 19.11).

In terms of conflict of interest, attention has been focused on whether financial ties to one drug company are associated with favorable results or conclusions. These ties have been questioned both as it relates to authors but also to journals. This has led the Cochrane Collaboration to put out a statement of its current policy that states "the sponsorship of a Cochrane review by any commercial source or sources…is prohibited" [28]. However, this area has been dominated by perceptions and not necessarily fact. Yank et al. [29] attempted to study financial ties by evaluating 124 meta-analyses that evaluated the effects of antihypertensive drugs in adults that compared any comparator on clinical endpoints. They concluded that "meta-analyses on antihypertensive drugs and with financial ties to one drug company are not associated with favorable results but are associated with favorable conclusions" (a so-called "spin" on the interpretation of the results) and that this discordance was not apparent in studies supported by non-profit groups. In an effort to address the financial conflict of interest and the impact that it has on the results of trials, Aneja et al. studied this question with respect to major cardiovascular trials. In their analysis they found that "self declared financial conflict of interest and source of funding do not seem to impact outcomes…" [30] and that a sub-analysis based upon the type of funding, or the selection of a surrogate over a clinical endpoint also did not seem to increase the likelihood of favorable trial results. In an accompanying editorial by Califf [30] some limitations of Aneja's results was pointed out (e.g. three major journals were selected and how representative these journals were compared to all the literature was pointed out, along with the fact that self-reported financial conflict of interest could be inaccurate).

One major concern about conflicts of interest revolves around the development of clinical practice guidelines, since these guidelines are being increasingly used in malpractice cases and for forming the basis of many pay-for-performance initiatives. For example a study published in 2011 reported that more than half of the participants involved in writing recent American College of Cardiology/American Heart Association clinical practice guidelines reported some financial conflict of interest [31]. Rochan et al. developed a financial conflict of interest checklist for clinical research studies and invited comments, but there is still wide variation in requirements [32]. Controversy even exists about the term "conflict" which Weber points out "*…almost implies that in order to receive the funding to do the research, the physician had to do something that had an adversarial or negative impact on the patients he was caring for.*" [33] *Indeed, Stossel states in that same article that "medicine is incomparably better than when I started out practicing about 40 years ago," it is not because doctors are now somehow more ethical or have been more heavily regulated — rather, it is because of the products that they have developed and gotten through their collaborations with industry.*

Another trend that is occurring with regard to manuscript publication is the increased frequency of open-access journals. Part of the justification for open access journals is the flawed peer-review process. Horton (Richard Horton, FRCP FMedSci, editor-in-chief of The Lancet) has opined, "… *we know that the system of peer review is biased, unjust, unaccountable, incomplete, easily fixed, often insulting, usually ignorant, occasionally foolish, and frequently wrong.*" We also know that the agreement between reviewers is often low, reviewers miss many mistakes, and reviewers can be biased against certain institutions and work that disagrees with what they have published. Peer review has resulted in the rejection of at least two papers that ultimately led to Nobel prizes; and that part of the reason for this is that there is little reward for the time spent in peer review, either monetarily of towards promotion.

Open access journals are scholarly journals that are available online to the reader "without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself" [34, 35]. Open access got its start about a decade ago and quickly won widespread acclaim with the advent of well-regarded, peer-reviewed journals like those published by the Public Library of Science, known as PLoS. Such articles were listed in databases like PubMed, which is maintained by the National Library of Medicine, and selected for their quality.

Some open access journals are subsidized, and some require payment on behalf of the author. Subsidized journals are financed by an academic institution, learned society or a government information center; those requiring payment are typically financed by money made available to researchers for the purpose from a public or private funding agency, as part of a research grant. There have also been several modifications of open-access journals that have considerably different natures: hybrid open-access journals and delayed open-access journals.

Open-access journals may be considered as:

- Journals entirely open access
- Journals with research articles open access (hybrid open-access journals)

- Journals with some research articles open access (hybrid open-access journals)
- Journals with some articles open access and the other delayed access
- Journals with delayed open access (delayed open-access journals)
- Journals permitting self-archiving of articles

Advantages and disadvantages of open access journals are the subject of much discussion amongst scholars and publishers. A few obvious advantages of open access journals include the free access to scientific papers regardless of affiliation with a subscribing library, lower costs for research in academia and industry, in addition to improved access for the general public and higher citation rates for the author. The argument for open access is that peer review has many problems by itself, and it has become increasingly difficult to find qualified peer reviewers willing to spend uncompensated time for that task. For open access journals, it is expected that the reader will act as the peer reviewer, but some researchers are now raising the alarm about what they see as the proliferation of online journals that will print seemingly anything for a fee. They warn that non-experts doing online research will have trouble distinguishing credible research from junk. In fact Jeffrey Beall, a librarian at Auraria Library, University of Colorado Denver, in Denver, Colorado, has been posting frequently updated lists of potential predatory open access journals [36] and Nissan [37] cites an example reported in the New Science Magazine of a hoax designed to test the legitimacy of a certain publisher.

Another consideration in manuscript preparation is the expense of publishing, thus, manuscripts must be as brief as possible. And many journals are moving toward "open access" publications where the cost of the publication is borne by the author. To emphasize the brevity that manuscripts must strive for, a rather humorous exchange has been published in;

THE JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1974, 7, 497 NUMBER 3, entitled THE UNSUCCESSFUL SELF-TREATMENT OF A CASE OF "WRITER'S BLOCK" by DENNIS UPPER, VETERANS ADMINISTRATION HOSPITAL, BROCKTON, MASSACHUSETTS

Abstract, None
Introduction, none
Methods and Results, None
Discussion, Blank
References, 0
Portions of this paper were not presented at the 81st Annual American Psychological Association Convention, Montreal, Canada, August 30, 1973. Reprints may be obtained from Dennis Upper, Behavior Therapy Unit, Veterans Administration Hospital, Brockton, Massachusetts 02401.
Received 25 October 1973. (Published without revision.)
COMMENTS BY REVIEWER A

I have studied this manuscript very carefully with lemon juice and X-rays and have not detected a single flaw in either design or writing style. I suggest it be published without revision. Clearly it is the most concise manuscript I have ever seen-yet it contains sufficient detail to allow other investigators to replicate Dr. Upper's failure. In comparison with the other manuscripts I get from you containing all that complicated detail, this one was a pleasure to examine.

Surely we can find a place for this paper in the Journal-perhaps on the edge of a blank page.

A follow-up manuscript was published some years later in the same Journal (JOURNAL OF APPLIED BEHAVIOR ANALYSIS 2007, 40, 773 NUMBER 4 (WINTER 2007)) entitled A MULTISITE CROSS-CULTURAL REPLICATION OF UPPER'S (1974) UNSUCCESSFUL SELF-TREATMENT OF WRITER'S BLOCK by ROBERT DIDDEN RADBOUD UNIVERSITY NIJMEGEN JEFF SIGAFOOS UNIVERSITY OF TASMANIA MARK F. O'REILLY UNIVERSITY OF TEXAS AT AUSTINGIULIO E. LANCIONI UNIVERSITY OF BARI PETER STURMEY QUEENS COLLEGE, CITY UNIVERSITY OF NEW YORK

ABSTRACT: None
INTRODUCTION: None
METHODS and RESULTS: None
DISCUSSION: None
Reviewers Comments

The Consistency Between the Findings of This Multisite Cross-cultural Replication by Didden, Sigafoos, O'Reilly, Lancioni, and Sturmey and those reported in Upper's classic paper on writer's block (Upper, 1974) are remarkable and serve to substantially extend the generality of Upper's findings.

The consistency between the editorial opinion of the action editor, Linda LeBlanc, whose reviewer comments are enclosed verbatim parenthetically here ( ) and this paper is equally remarkable.

This kind of symmetry is rare in any science and particularly rare in behavior analysis, and because of it I was compelled to accept the Didden et al. paper without revision. I did not change one word, and this is a first in my tenure as editor. Another virtue of the paper is its awe-inspiring brevity. It is my hope that it will one day serve as the model for Brief Reports in JABA.

Preparation of this article was supported by a grant of $2.50 from the first author's personal funds. We hope to submit a version of this paper at the next international conference in St. Tropez. Received July 2, 2007 Final acceptance July 5, 2007.

References

1. van Ekelenburg H. The art of writing good research proposals. Sci Prog. 2010;93:429–42.
2. Fosbol EL. Major medical (Meetings). CardioSource WorldNews. 2012;47.
3. Winnik S, Raptis DA, Walker JH, Hasun M, Speer T, Clavien PA, et al. From abstract to impact in cardiovascular research: factors predicting publication and citation. Eur Heart J. 2012;33:3034–45. PMC3530902.

4. Krzyzanowska MK, Pintilie M, Tannock IF. Factors associated with failure to publish large randomized trials presented at an oncology meeting. JAMA. 2003;290:495–501.
5. Krzyzanowska MK, Pintilie M, Brezden-Masley C, Dent R, Tannock IF. Quality of abstracts describing randomized trials in the proceedings of American Society of Clinical Oncology meetings: guidelines for improved reporting. J Clin Oncol. 2004;22:1993–9.
6. Want to be taken seriously? Become a better writer. In MSN. www.msn.com. Accessed 25 Feb 2013.
7. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS Med. 2010;7:e1000251.
8. Quoteland.com. http://www.netrax.net/~rarebook/s971030.htm. Accessed 12 Feb 2013.
9. Welch HG. Preparing manuscripts for submission to medical journals: the paper trail. Eff Clin Pract. 1999;2:131–7.
10. Altman DG. Comparability of randomised groups. Statistician. 1985;34:125–36.
11. Branson RD. Anatomy of a research paper. Respir Care. 2004;49:1222–8.
12. Grammarly: the world's best grammar checker. http://grammerly.com. Accessed 12 Feb 2013.
13. Accad M. Statistics and the rise of medical fortunetellers. Tex Heart Inst J. 2009;36:508–9. PMC2801944.
14. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4:e296. PMC2020495.
15. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6:e1000100. PMC2707010.
16. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283:2008–12.
17. Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, Von Elm E, et al. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. PLoS Med. 2009;6:e22. PMC2634792.
18. Wislar JS, Flanagin A, Fontanarosa PB, Deangelis CD. Honorary and ghost authorship in high impact biomedical journals: a cross sectional survey. Br Med J. 2011;343:d6128.
19. Mowatt G, Shirran L, Grimshaw JM, Rennie D, Flanagin A, Yank V, et al. Prevalence of honorary and ghost authorship in Cochrane reviews. JAMA. 2002;287:2769–71.
20. Greenland P, Fontanarosa PB. Ending honorary authorship. Science. 2012;337:1019. doi:10.1126/science.1224988.
21. Wilhite AW, Fong EA. Scientific publications. Coercive citation in academic publishing. Science. 2012;335:542–3. doi: 10.1126/science.1212540.
22. Franck G. Essays on science and society. Scientific communication – a vanity fair? Science. 1999;286:53–5.
23. Hirsch LJ. Conflicts of interest, authorship, and disclosures in industry-related scientific publications: the tort bar and editorial oversight of medical journals. Mayo Clin Proc. 2009;84:811–21.
24. Booth CM, Cescon DW, Wang L, Tannock IF, Krzyzanowska MK. Evolution of the randomized controlled trial in oncology over three decades. J Clin Oncol. 2008;26:5458–64. PMC2651075.
25. Rose SL, Krzyzanowska MK, Joffe S. Relationships between authorship contributions and authors' industry financial ties among oncology clinical trials. J Clin Oncol. 2010;28:1316–21. PMC3040064.
26. Chalmers TC, Smith Jr H, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials. 1981;2:31–49.
27. Kaiser KA, Cofield SS, Fontaine KR, et al. Is funding source related to study reporting quality in obesity or nutrition randomized control trials in top-tier medical journals? Int J Obes. 2012;36:977–81.
28. Commercial sponsorship and the Cochrane Collaboration: the Cochrane Collaboration policy on commercial sponsorship. Revised April 2006. Accessed at www.cochraneorganization/docs/commercial/sponsorship

29. Yank V, Rennie D, Bero LA. Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. Br Med J. 2007;335:1202–5.
30. Califf RM. Conflicting information about conflict of interest. J Am Coll Cardiol. 2013;61:1137–43. doi:10.1016/j.jacc.2012.12.030.
31. Blum JA, Freeman K, Dart RC, Cooper RJ. Requirements and definitions in conflict of interest policies of medical journals. JAMA. 2009;302:2230–4. doi:10.1001/jama.2009.1669.
32. Rochon PA, Hoey J, Chan A-W, Ferris LE, Lexchin J, Kalkar SR, et al. Financial Conflicts of Interest checklist 2010 for clinical research studies. Open Med. 2010;4:e69–91. PMC3116675.
33. Weber M. Conflict of interest: an outdated phrase for physician-industry relationships. Orthopedics Today. 2010. Accessed at: http://www.healio.com/orthopedics/business-of-orthopedics/news/print/orthopedics-today/%7Bd5f9514f-4796-47ce-a8c6-a02ef9b87a5f%7D/conflict-of-interest-an-outdated-phrase-for-physician-industry-relationships
34. Open Access Overview. Retrieved on November 11, 2012. Accessed at http://www.planta.cn/forum/files_planta/what_is_open_accessan_overview_2004_162.pdf
35. No-fee open-access journals. 2006. Accessed at http://dash.harvard.edu/bitstream/handle/1/4552050/suber_nofee.htm?sequence=1
36. Scholarly Open Access: critical analysis of scholarly open-access publishing. Accessed 2013, at www.scholarlyoa.com
37. Nissan S. Predatory Publishers; Authors Beware. CardioSource WorldNews. 2013;42.

Chapter 20
The Media and Clinical Research

Stephen P. Glasser

"Media is a word that has come to mean bad journalism"
http://thinkexist.com/search/

Abstract The news media are an increasingly important source of information about new medical treatments. The media can be persuasive, pervasive, and can influence health care beliefs and behaviors. This chapter briefly addresses the maturation process of medical controversy, discusses some of the reasons for the "tension" that develops between scientists and the media, and hopefully allows the reader when they are asked to discuss their research findings, to develop some strategies for dealing with the media.

Keywords Media in clinical research • Medical controversies • Embargo rule • Academic health center and the media

The media (whether we like it or not) is playing an increasing role in helping or confounding the transmission of knowledge to patients. The news media are an increasingly important source of information about new medical treatments. The media can be persuasive, pervasive, and can influence health care beliefs and behaviors [1]. Caspermeyer et al. investigated nine large newspapers to determine how often the coverage of neurological illness contained errors and stigmatizing language [2]. They determined that medical errors occurred in 20 % and stigmatizing language in 21 % of the articles evaluated. In another report, seven stories regarding three preventative treatments (cholesterol, osteoporosis, and aspirin) were analyzed [3]. Of those media reports, 40 % did not report benefits quantitatively; of those that did, 83 % reported relative (not absolute) benefits only, while 98 % reported potential harm.

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_20,

The Natural History of Medical Controversies (Table 20.1)

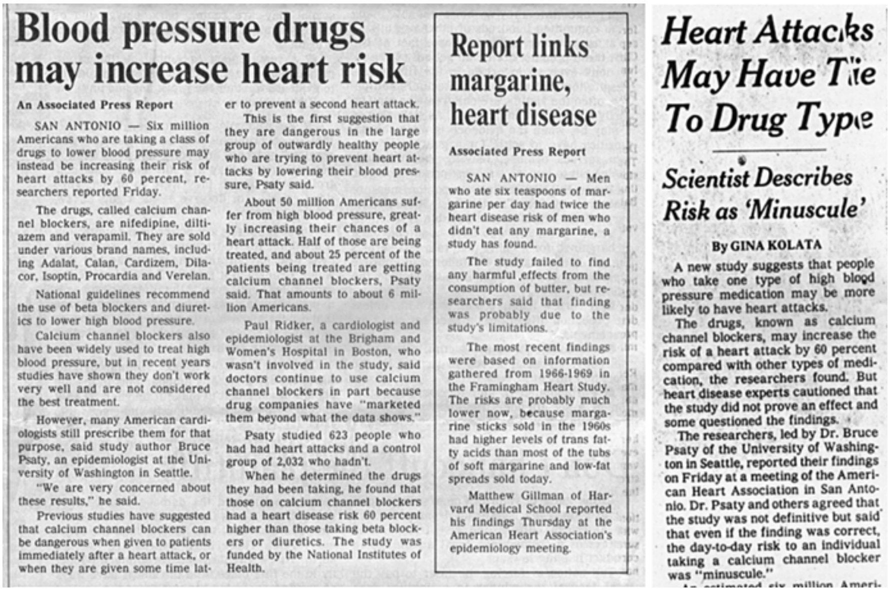
In 1997 Weber reviewed the "natural history" of reports on medical controversies (approximately a 10 year evolution) which I believe are instructional [4]. The first phase in the natural history of media reports about medical innovations, he entitled the Genesis Phase. During the Genesis Phase new information is identified. The next phase in the natural history of media reporting is the Development Phase, where questions of safety and/or efficacy about the innovation arise; print and broadcast publicize the debate; and, complex issues tend to be oversimplified and/or sensationalized. This is followed by the Maturation Phase where more data and studies become available, but public interest by this time tends to be waning and media coverage is less intense. Finally, there is the Resolution Phase where objective re-evaluations are published, and a more fair-balance of the pros and cons of the innovation are presented (Table 20.1). Weber presents two examples of this natural evolution process: the silicone gel breast implant; and, the calcium channel blocker (CCB) controversies, the latter of which is discussed below.

The genesis of the CCB controversy began in 1995 when Psaty et al. presented a Case Control Study from a single center suggesting that short-acting nifedipine could harm patients treated for hypertension (specifically they reported an increased risk of myocardial infarction) [5]. The RR for harm was reported as 1.6. The Development Phase was evident after the American heart Association published a press release that was hyped by the media. Many who were treating patients with hypertension at that time will recall being inundated with telephone calls from concerned patients. Examples of the news reports are shown in Fig. 20.1.

The CCB controversy that arose was followed by a meta-analysis (see Chap. 10) of 16 studies also suggesting the same harm [6]. Subsequently, all CCBs were said to be harmful and furthermore were additionally said to be associated with cancer and GI bleeding [5–7]. During the Maturation Phase of this controversy, the FDA and NIH reviewed the CCB data and gave them a clean bill of health (with the exception of short-acting CCBs). Reanalysis of the data began to show the flaws in the methodology of studies impugning the CCBs. The methodological flaws included selection bias and prescription bias, that is, sicker patients were more likely to be given CCBs. In the Resolution Phase (8–10 years after the controversy began), the CCB controversy was "put to rest" most recently by ALLHAT [8]. It should be noted that during this process another issue surfaced relative to the Multicenter Isradipine Diuretic Atherosclerosis Study (MIDAS), a large multicenter study that compared the effects of isradipine (a short-acting CCB) compared

Table 20.1 The 10 year "natural history" of medical controversies

Genesis phase	New information is identified
Development phase	Questions of safety and/or efficacy arise
Maturation phase	More data and studies become available; interest by public and media wanes
Resolution phase	Objective re-evaluations are published



Blood pressure drugs may increase heart risk

An Associated Press Report

SAN ANTONIO — Six million Americans who are taking a class of drugs to lower blood pressure may instead be increasing their risk of heart attacks by 60 percent, researchers reported Friday.

The drugs, called calcium channel blockers, are nifedipine, diltiazem and verapamil. They are sold under various brand names, including Adalat, Calan, Cardizem, Dilacor, Isoptin, Procardia and Verelan.

National guidelines recommend the use of beta blockers and diuretics to lower high blood pressure.

Calcium channel blockers also have been widely used to treat high blood pressure, but in recent years studies have shown they don't work very well and are not considered the best treatment.

However, many American cardiologists still prescribe them for that purpose, said study author Bruce Psaty, an epidemiologist at the University of Washington in Seattle.

"We are very concerned about these results," he said.

Previous studies have suggested that calcium channel blockers can be dangerous when given to patients immediately after a heart attack, or when they are given some time later to prevent a second heart attack.

This is the first suggestion that they are dangerous in the large group of outwardly healthy people who are trying to prevent heart attacks by lowering their blood pressure, Psaty said.

About 50 million Americans suffer from high blood pressure, greatly increasing their chances of a heart attack. Half of those are being treated, and about 25 percent of the patients being treated are getting calcium channel blockers, Psaty said. That amounts to about 6 million Americans.

Paul Ridker, a cardiologist and epidemiologist at the Brigham and Women's Hospital in Boston, who wasn't involved in the study, said doctors continue to use calcium channel blockers in part because drug companies have "marketed them beyond what the data shows."

Psaty studied 623 people who had had heart attacks and a control group of 2,032 who hadn't.

When he determined the drugs they had been taking, he found that those on calcium channel blockers had a heart disease risk 60 percent higher than those taking beta blockers or diuretics. The study was funded by the National Institutes of Health.

Report links margarine, heart disease

Associated Press Report

SAN ANTONIO — Men who ate six teaspoons of margarine per day had twice the heart disease risk of men who didn't eat any margarine, a study has found.

The study failed to find any harmful ,effects from the consumption of butter, but researchers said that finding was probably due to the study's limitations.

The most recent findings were based on information gathered from 1966-1969 in the Framingham Heart Study. The risks are probably much lower now, because margarine sticks sold in the 1960s had higher levels of trans fatty acids than most of the tubs of soft margarine and low-fat spreads sold today.

Matthew Gillman of Harvard Medical School reported his findings Thursday at the American Heart Association's epidemiology meeting.

Heart Attacks May Have Tie To Drug Type

Scientist Describes Risk as 'Minuscule'

By GINA KOLATA

A new study suggests that people who take one type of high blood pressure medication may be more likely to have heart attacks.

The drugs, known as calcium channel blockers, may increase the risk of a heart attack by 60 percent compared with other types of medication, the researchers found. But heart disease experts cautioned that the study did not prove an effect and some questioned the findings.

The researchers, led by Dr. Bruce Psaty of the University of Washington in Seattle, reported their findings on Friday at a meeting of the American Heart Association in San Antonio. Dr. Psaty and others agreed that the study was not definitive but said that even if the finding was correct, the day-to-day risk to an individual taking a calcium channel blocker was "minuscule."

An estimated six million Ameri-

Fig. 20.1 Two examples of media reports on the CCB controversy

to the diuretic hydrochlorothiazide on the course of carotid artery disease in hypertensive patients [9]. The investigators found that the progression of carotid atherosclerosis did not differ between the two treatment groups, but that there was an increased incidence of vascular events in patients treated with the CCB. A side issue in this study was the withdrawal of some of the investigators from the manuscript preparation due to what they perceived as "undue influence" exerted by the sponsor of the study (See Chap. 19 and conflict of interest). Needless to say, this resulted in some interesting media reporting such as a headline that said "a high-tension drug study has been reported".

The "Tension" Between Scientists and the Media

Why the media publicized the CCB controversy and deemed it newsworthy while another controversy is not so publicized seems to be a mystery to most readers and listeners. In great part the publicizing of such studies depends upon what the media editors think will have "headline potential". As Semir noted, "…news of killer bacteria, exterminating viruses, and miraculous therapies tend to have greater appeal because such stories compete with murders, rapes, ecologic catastrophes, and declarations from famous people…" [10]. In fact, this author had a personal experience following publication of 13 subjects who underwent a roll-a-coaster ride [11]. The heart rate response (by ambulatory ECG monitoring) was quite impressive; but, let's face

Table 20.2 Questions cited by one science media reporter

Was the study large enough
Was the study fair
Who paid for the study
Who was the control group
Were volunteers randomly assigned
Was there appropriate blinding

it, 13 healthy subjects with no adverse outcomes? Yet this became a story for national media attention, probably because there had been a few recent deaths on similar rides throughout the country. Marilyn Chase reported in the Wall Street Journal ways of putting hyped study results under the microscope [12]. Every week, she noted, medical science makes headlines with a promising new study or "cure", and it is "often hard to tell ephemeral findings from epochal breakthroughs-especially when distilled into a few paragraphs or sound bites spiced with hype" [12]. Interestingly, she cites a number of questions that need to be addressed in media reports, questions that should sound familiar from reading chapters in this book, regarding clinical trial methodology. Some of the questions Chase cited were: Was the study large enough to make it significant? Was the study fair i.e. were the two groups equally matched? Who paid for the study? Who was the control group? Were volunteers randomly assigned? Was there appropriate blinding? (Table 20.2)

Deary et al. report their media experience with a study that had been reported in Lancet [13]. The Lancet report concluded that women with more submissiveness were less likely to have myocardial infarction compared to those women who were less submissive. The Lancet publication was under embargo (a topic to be discussed shortly); however, a newspaper ran the story prematurely under the headline "put down that rolling pin, darling, it's bad for your heart". Other headlines included "do as you're told girls…and live to be old", "stay home and you'll live longer", "do what hubby says and you will live longer", and "meekness is good for a women's heart…" The authors further note that one phone interview included questions like: "So these feminists are all barking up the wrong tree?" and, "Should women be getting back to the kitchen sink?" Of course, these questions did not accurately represent what the study in fact showed, and I recommend reading Deary's editorial, as it should be instructive to all researchers interested in communicating their studies results.

The Importance of the Media in Providing Health Information

The importance of the media in providing the public with health information should not be underestimated. Timothy Johnson (in the 108th Shattuck Lecture) noted a survey in which 75 % of the respondents said they pay either a great deal or moderate amount of attention to the medical and health news reported by the media; and, 58 % said that they have changed their behavior or have taken some type of action based upon what was reported (read, seen, or heard) [14]. Thus, the role of the clinical researcher in providing news to the media is important. Some basic tenants for the researcher to follow are: be certain you are the best person to provide the media with the necessary information; do not digress – start with your main conclusion first and then do not wander; consider the 2–3 points that are important about one's study, and keep returning to those points; do not become defensive or argumentative; and, be concise – particularly with television interviews. As an example of the above let us assume that you have hypothetically just published a study on the benefits of a new drug and the interview proceeds with a question such as "what were your primary findings?" Having briefly discussed the outcomes with great pride, the reporter than asks "but doctor weren't there three deaths in your study and do you really think it was ethical to perform such a trial?" The response by most of the uninitiated would go something like this- "yes there were three deaths, but in this population we expected there to be deaths, and blah blah blah". In general it is best not to repeat the negative, and the answer perhaps could have been better shaped with something like "the important thing is that we found a significant overall benefit of our new drug treatment, and this was in a very sick population. In addition we did everything possible to protect the safety of our patients." Many might remember the very funny interview in the Bob Newhart comedy television series, when off camera a very pleasant reporter pumped up Newhart's ego, and when they went live totally blind-sided him with embarrassing and demeaning questions such as "since psychologists hardly ever cure anyone, don't you think the fees that you charge them are outrageous?". In actuality, this type of blind-siding is rare with health reporting, the reporter is generally your colleague, and is attempting (with their limited knowledge) to impart accurate information, but being prepared for that occasional problem is not a bad idea.

The Medias Control of Information (The Embargo Rule)

Perhaps the most important issue that results in researcher-media conflicts is the long struggle over the "Ingelfinger rule" since it involves the control of information, a control the media despises. The pressure to be the first or to be able to claim to be the exclusive report of a story results in significant tension when they are asked to hold (embargo) a story until it is published in a scientific journal.

Scientists also expect that they are the ones to control the flow of information, and view the media as but a pipeline to inform the public about recent discoveries [1]. Most journalists, however, do not view themselves merely as a spokesperson for the scientist, but rather they view their role as raising probing questions about the research. In fact, both scientists and journalists are committed to communicating accurate information, but the media aims for brevity, readability, simplicity; and, are usually pressured by time constraints; whereas the scientist has been working on the research that is being reported for years, are interested in precautionary qualifications, and are aware that their scientific readership can assimilate the nuances of their research [3].

The Academic Medical Centers Role in Promoting Health Research

In Academic Medical Centers the investigator interaction with the media is frequently channeled through a media relations or public relations office that provides some insulation for the investigator. Woloshin et al. noted that “medical journalism is often criticized for what reporters cover…and how they cover it” and also note that the tension between scientists and journalists exists because scientists want to promote the truth, the media just wants to sell newspapers [15]. In order to assess the role that academic medical centers play in the release of scientific information (with the presumption that their releases would be measured and unexaggerated), the authors examined press releases from academic medical centers in a systematic fashion. The details of the study can be found by referring directly to the cited reference, but what they found was that there were a mean of 49 press releases annually. That 44 % promoted animal or laboratory research, of which 74 % claimed human health relevance; and, that among 95 press releases about human research, 23 % omitted study size, 34 % failed to quantify results, and only 17 % promoted studies with the strongest study designs. Furthermore they found that 40 % reported on uncontrolled interventions, small sample size studies, surrogate outcome studies, and yet 585 lacked the relevant cautions. Among the recommendations suggested by the authors they listed that they should include basic facts and explicit cautions, that investigators should forego responding to requests for releases of studies with obvious limitations, taking care to temper their tone if they do respond, and that journalists have the opportunities to acquire skills through a number of programs and workshops available to them.

In summary, the media is playing an increasing role in the reporting of health news. Most health reporters are attempting to write a credible and accurate story. The enduring tensions between medicine and the media are largely due to the different perspectives between researchers and journalists. As Nelkin noted, “these tensions arise because of perceived differences in defining science news, conflicts over styles of science reporting, and most of all disagreement about the role of the media” [16]. It is incumbent upon the researcher, if they are going to accept a media interview, to know how to present clear concise answers to question about their research.

References

1. Fishman JM, Casarett D. Mass media and medicine: when the most trusted media mislead. Mayo Clin Proc. 2006;81:291–3.
2. Caspermeyer JJ, Sylvester EJ, Drazkowski JF, Watson GL, Sirven JI. Evaluation of stigmatizing language and medical errors in neurology coverage by US newspapers. Mayo Clin Proc. 2006;81:300–6.
3. Moynihan R, Bero L, Ross-Degnan D, Henry D, Lee K, Watkins J, et al. Coverage by the news media of the benefits and risks of medications. N Engl J Med. 2000;342:1645–50.

4. Psaty BM, Heckbert SR, Koepsell TD, Siscovick DS, Raqhunathan TE, Weiss NS, et al. The risk of myocardial infarction associated with antihypertensive drug therapies. JAMA. 1995;274:620–5.
5. Weber MA. The natural history of medical controversy. Consultant. 1997.
6. Furberg C, Psaty B, Meyer J. Nifedipine. Dose-related increase in mortality in patients with coronary heart disease. Circulation. 1995;92:1326–31.
7. Jick H. Calcium-channel blockers and risk of cancer. Lancet. 1997;349:1699–700.
8. Pahor M, Guralnik J, Furberg CD. Risk of gastrointestinal hemorrhage with calcium antagonists in hypertensive patients over 67. Lancet. 1996;347:1061–6.
9. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group, The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). JAMA. 2002;288:2981–97.
10. Borhani NO, Mercuri M, Borhani PA, Buckalew VM, Canossa-Terris M, Carr AA, et al. Final outcome results of the Multicenter Isradipine Diuretic Atherosclerosis Study (MIDAS). A randomized controlled trial. JAMA. 1996;276:785–91.
11. de Semir V. What is newsworthy? Lancet. 1996;347:1163–6.
12. Glasser SP, Clark PI, Spoto E. Heart rate response to “fright stress”. Heart Lung J Acute Crit Care. 1978;7:1006–10.
13. Chase M. How to put hyped study results under a microscope. Wall Street J. 1995;16:B-1.
14. Deary IJ, Whiteman MC, Fowkes FG. Medical research and the popular media. Lancet. 1998;351:1726–7.
15. Johnson T. Shattuck lecture–medicine and the media. N Engl J Med. 1998;339:87–92.
16. Nelkin D. An uneasy relationship: the tensions between medicine and the media. Lancet. 1996;347:1600–3.

Chapter 21
Mentoring and Advising

Stephen P. Glasser and Edward W. Hook III

> *"Advice is like mushrooms. The wrong kind can prove fatal."*
>
> –unknown

Abstract Mentorship refers to the development of a relationship between a more experienced individual (the mentor) with a less experienced individual (the mentee or protégé). The role and expectations of the mentor in the development of the junior faculty member's academic relationship is extremely important. As such, this chapter discusses the expectations of the mentor, mentee, and the mentor-mentee relationship.

Keywords Mentorship • Mentoring guidelines • Advising

Mentoring vs. Advising

Mentorship refers to the development of an ongoing, advisory relationship between a more experienced individual (the mentor) with a less experienced individual (the mentee or protégé). Historically, mentorship goes back to ancient Greek and Hindu times and the word itself was inspired by the character of Mentor in Homer's Odyssey. Today, the definition of mentor continues to encompass 'a trusted

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

E.W. Hook III, M.D.
Division of Infectious Diseases, Epidemiology and Microbiology, STD Control Program,
Jefferson County Department of Health, University of Alabama at Birmingham,
Birmingham, AL, USA

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_21,

counselor or guide', and a 'wise, loyal advisor or coach.' True mentoring however is more than just answering occasional questions or providing *ad hoc* help. It is about an ongoing relationship of learning, dialog, and challenge. "Mentoring" is a process that always involves communication and is relationship based, but its precise definition is elusive. One of the many definitions that have been proposed, is: mentoring is a process for the informal transmission of knowledge, social capital, and the psychosocial support perceived by the recipient as relevant to work, career, or professional development; "mentoring entails informal communication, usually face-to-face and during a sustained period of time, between a person who is perceived to have greater relevant knowledge, wisdom, or experience (the mentor) and a person who is perceived to have less (the protégé)" [1].

Mentoring in the research sense developed mostly in the basic science laboratories, where an experienced researcher would literally take a junior person 'under their wing' and would help them develop research independence. This concept has been adopted and encouraged by the NIH through its K23 and K24 programs which, in turn, serve as templates for career development programs supported by other organizations. The problem has always been, that there is little in the way of formal training in how to be a good mentor, and there is usually little external reward for the time spent in mentoring.

In academic settings, mentoring and academic advising are frequently used synonymously, but we view advising as a lesser responsibility than mentoring. One can over-simplistically say that advising is an 'event' while mentoring is a 'process'. A mentor has both a professional and personal relationship with the mentee while an advisor, in general, does not have the same sort of personal relationship. Also, mentoring is more dynamic, in that there is a distinct, evolutionary change over time.

Although there is no single formula for good mentoring, most would agree that a good mentor is approachable and available, and this is where good mentoring too often comes up short, since in a busy academician's life (who has multiple demands including their own requirements for promotion, research grants, manuscripts, etc.); little academic reward is provided for mentoring. It is for this reason that, although perhaps more empathetic with the role of the mentee, junior faculty are often ill-equipped to serve as mentors. Factors militating against effective mentorship by junior faculty include an (appropriate) emphasis on one's own career advancement, limited resources to devote to the mentee, and limited opportunities to promote the mentee's career by virtue of limited personal recognition as a result of being early in one's career. Students, for their part, must recognize the professional pressures and time constraints faced by their mentors, but still must insist on obtaining adequate time and availability from their mentors, or be willing to change who their mentor is. Much misunderstanding can be circumvented with a well-intentioned discussion about these issues prior to choosing a given mentor. As such, both the mentor and mentee should be clear about their respective expectations, have a clear agreed upon career development plan, with regular meetings a priority. On the one hand, the mentor cannot be to busy, otherwise they should not have accepted the responsibility, but the mentee cannot expect unlimited access.

Table 21.1 Five common techniques used by mentors

Accompanying: Committing in a caring way
Sowing: Laying the foundation even if the mentee does not yet understand its importance
Catalyzing: Plunging the mentee into a new way of thinking
Showing: Making something understandable
Harvesting: What have you learned, and how useful is it

From: Aubrey and Cohen [3]
This material is reproduced with permission of John Wiley & Sons, Inc.

In some instances the use of a "mentoring contract" which both the mentor and mentee work together to delineate the goals and structure of the relationship in writing can provide the clarity of purpose that is the foundation of most successful mentoring relationships.

Prior research has suggested that mentorship in academic health science centers has an important influence on productivity and personal development [2]. But, most programs have been modest in scope. Feldman et al. did analyze the baseline variables prior to instituting a structured comprehensive mentoring program at one institution in order to assess the characteristics associated with having a mentor along with the content of mentor-mentee interactions. More than half the respondents to a survey (with a 56 % response rate) stated that they had a mentor, and that there were no differences in having a mentor based upon gender or ethnicity. Having a mentor was associated with greater satisfaction with time allocation at work, and reported that discussions of funding, manuscript preparation, promotion and tenure were among the most important topics.

A 1995 study of mentoring techniques most commonly used in business [3] found that the five most commonly used techniques among mentors were (Table 21.1):

1. *Accompanying:* making a commitment in a caring way, which involves taking part in the learning process side-by-side with the learner.
2. *Sowing:* mentors are often confronted with the difficulty of preparing the learner before he or she is ready to change. Sowing is necessary when you know that what you say may not be understood or even acceptable to learners at first but will make sense and have value to the mentee when the situation requires it.
3. *Catalyzing:* when change reaches a critical level of pressure, learning can escalate. Here the mentor chooses to plunge the learner right into change, provoking a different way of thinking, a change in identity or a re-ordering of values.
4. *Showing:* this is making something understandable, or using your own example to demonstrate a skill or activity. You show what you are talking about, you show by your own behavior.
5. *Harvesting:* here the mentor focuses on "picking the ripe fruit": it is usually used to create awareness of what was learned by experience and to draw conclusions. The key questions here are: "What have you learned?", "How useful is it?".

Guidelines for Faculty/Student Interactions

Faculty members often develop a close working relationship with students, especially advisees. Often a relationship is formed that provides benefits to both the faculty member and the student. Faculty should be cognizant of the power differential in these types of relationships and set appropriate boundaries. Although faculty members may not intend a favor or request to be an obligation, they should be aware that this may place some students in a difficult position. Some students are intimidated by faculty members and may not feel free to decline such requests [4]. It is recognized that many situations are ambiguous. Examples are of some of these ambiguous situations include:

- **Asking a student to drive you someplace, including the airport, home, or main campus**. Such a request does not fall under a student's duties. A situation when this may be acceptable is when the student has the same destination.
- **Asking student to work extra hours or late hours**. Students should be expected to work the hours they are paid for. Students may volunteer to put in extra hours to gain more experience (e.g. grant writing) or gain authorship on a paper or help meet a deadline – but these extra hours should not be an expectation.
- **Asking an advisee to housesit, take care of your children or pets, or help you move**. While some students may not mind house sitting, taking care of children or pets, or helping someone move, others may only agree to do this because they feel obligated or worry that saying no will somehow affect their relationship with the faculty member. To avoid this situation, faculty members may post a request for a sitter or mover for pay without any faculty names attached to the flyer – ensuring that respondents really want this job.

Advising

Expectations for advising vary between institutions but mainly in terms of frequency of meetings. It seems to these authors that minimal expectations should include (Table 21.2):

1. academic advisors should meet with their advisees at least twice per semester, but more often is preferable. These meetings should be scheduled, but there should also be opportunities for *ad hoc* meetings to deal with acute problems.
2. Academic advisors should respond in a timely manner to requests from advisees for meetings or responses by telephone or e-mail, even if this is to schedule the requested meeting.
3. Academic advisors should provide general guidance to students about coursework, fieldwork, project selection, and career planning.
4. Academic advisors should make students feel welcome to the Division.

Table 21.2 Advising expectations

Meet regularly: Scheduled not ad hoc
Respond in a timely manner to requests
Provide general guidance about course work, etc.
Be welcoming
Act as a contact person and direct to appropriate resources
Act as a resource for bureaucratic and political issues in the school
Balance over extending ones self with important opportunities

5. Academic advisors should act as a contact person for the student and help direct them to the appropriate resources in the Division given whatever issues or problems the students may have.
6. Academic advisors should act as a resource for the student when bureaucratic or political problems in the University, School or Division may be interfering with the student's effective progress toward his or her degree.
7. Although the advisors role is to help the advisee to not over-extend themselves, they should also help them see what an important opportunity is.

Advising may include a number of diverse activities such as procedural advising (e.g. should the student drop a course), academic advising (e.g. how satisfied are they with the program, career planning, selecting course work), and advising 'students' on the conduct of their research. Excellent advising requires a significant time commitment.

What are the mentor's responsibilities? They should find out what are the junior investigators career goals, determine how often formal meetings should take place, what the mentor's expectations are (this should be spelled out in terms of frequency of meetings, metrics, and outcomes), and devise the best way(s) to communicate (face to face, e-mail, telephone). The advisee also has responsibilities. They should take the lead in scheduling meetings, and contacting the advisor if there are problems. Finally, there should be clear expectations of what protected time will be provided for the mentee's career development. If this is not under the control of the mentor, the mentor should aid the mentee in establishing protected time with whoever the responsible person is. There are many pitfalls in the term 'protected time'. One of the most important is the denominator for calculating it. For example, is the % of protected time based upon a 40 h, 60 h, or 80 h-week. What other responsibilities will the mentee have (i.e. clinics, ward rotations, committee meetings, teaching, conferences etc.).

Mentoring Committees

With increasing emphasis on translational research as a career path, mentorship by committee has become more popular. This approach provides trainees with access to content experts in several different disciplines relevant to their career development and can be quite successful. There are several potential pitfalls to mentorship by committee as well. The benefits of a mentoring committee are maximized when the committee meets as a whole with the mentee to discuss plans and progress, not when the mentee is subjected to a series of individual meetings in which different mentors may differ in terms of their advice regarding prioritization and progress. A second common problem with mentoring by committee is the failure to identify a "primary" mentor who has major responsibility for advice to the trainees. When this does not occur and problems are encountered, a failure to take responsibility for the mentoring process can lead to confusion and misdirection for the mentee.

Effective mentorship has been shown to be essential for faculty career success and good mentoring relationships are more likely to result in the mentee remaining at an academic health center and be promoted more rapidly. Binkley and Brod point out that effective mentorship is also associated with greater career satisfaction, and better performance [5], Despite this, they note that at one large academic health center, the average prevalence of mentorship was 50 %.

K23 and K24 Awards (Figs. 21.1, 21.2, and 21.3)

The NIH has developed a number of Career Development Programs (K awards), in fact there are now 13 different awards available and these are dependent upon such factors as one's career stage and how they may interact with other NIH Awards. However, there are common elements of NIH career awards, such as specified levels of salary support, allocations for research/development costs, and award duration. In addition, entry-level awards require a mentor, and at least 75 % protected time for the awardees to spend on research and other career development activities. For non-mentored senior awards a 25–50 % time commitment is typically required. Eligibility for NIH awards requires a Doctoral Degree (generally), that the applicant be a US citizen, Non-Citizen National, or a Permanent Resident. Should the awardee change their Institution or Mentor prior approval of the NIH awarding component must be advised.

For most of the readers of this book, the K23 award is likely to be the most appropriate. The guidelines for K23 Awards include an application that includes information about the nature and extent of supervision that will occur during the award period (co-mentors must supply similar information), and there must also be a career development plan that incorporates a systematic approach towards obtaining the necessary skills necessary to become an independent researcher. This plan should include course work appropriate to the experience of the candidate. The mentor's research qualifications in the area of the project and the extent and quality of his/her proposed role in guiding and advising the mentee, as well as previous experience in mentoring are critical. The application must include the applicant's career goals and objectives with a detailed description of what the candidate wants to achieve following the completion of the award.

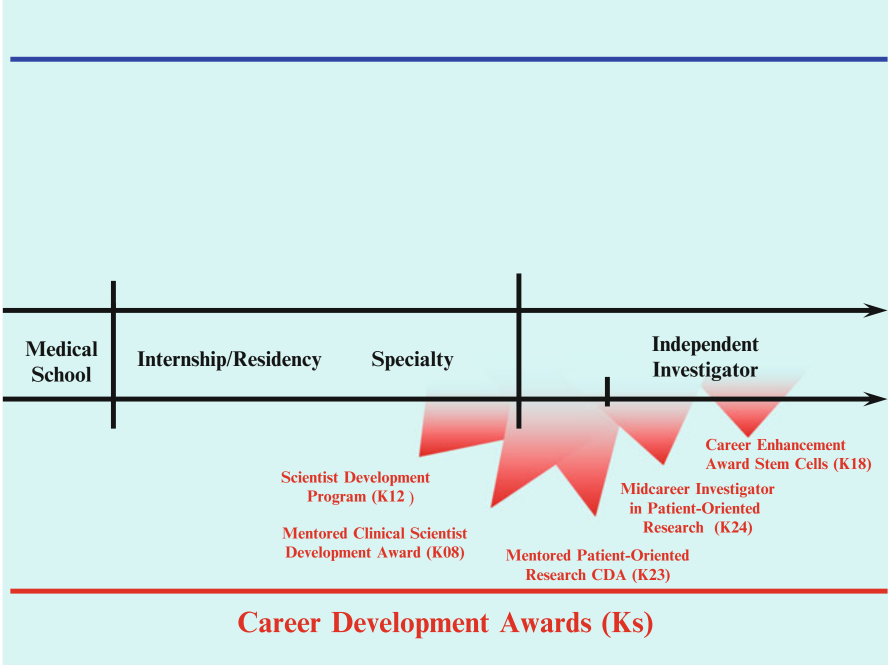
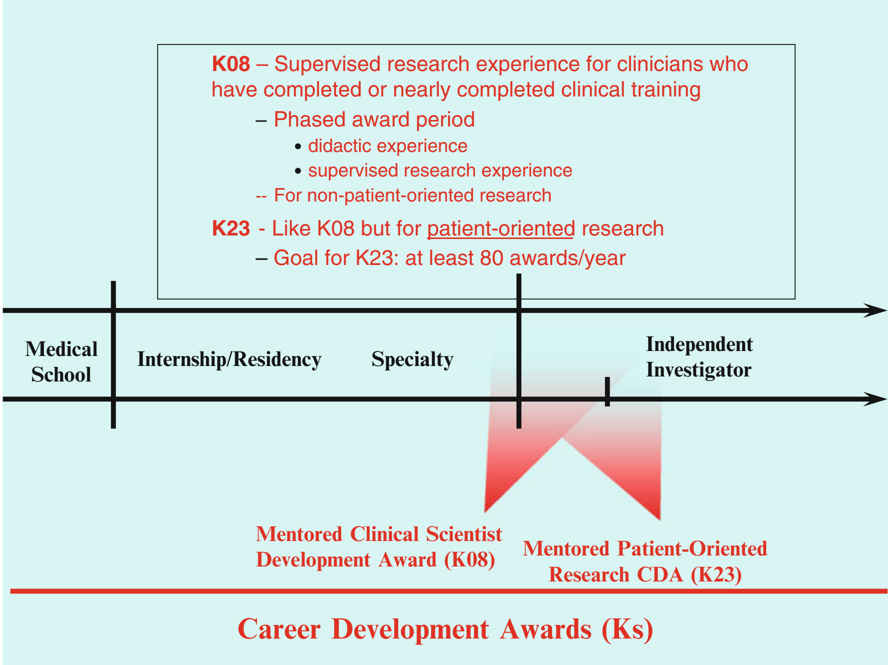

Medical School
Internship/Residency
Specialty
Independent Investigator
Scientist Development Program (K12)
Mentored Clinical Scientist Development Award (K08)
Mentored Patient-Oriented Research CDA (K23)
Midcareer Investigator in Patient-Oriented Research (K24)
Career Enhancement Award Stem Cells (K18)
Career Development Awards (Ks)

Fig. 21.1 The NIH career development awards (K awards)


K08 – Supervised research experience for clinicians who have completed or nearly completed clinical training

– Phased award period

- didactic experience
- supervised research experience

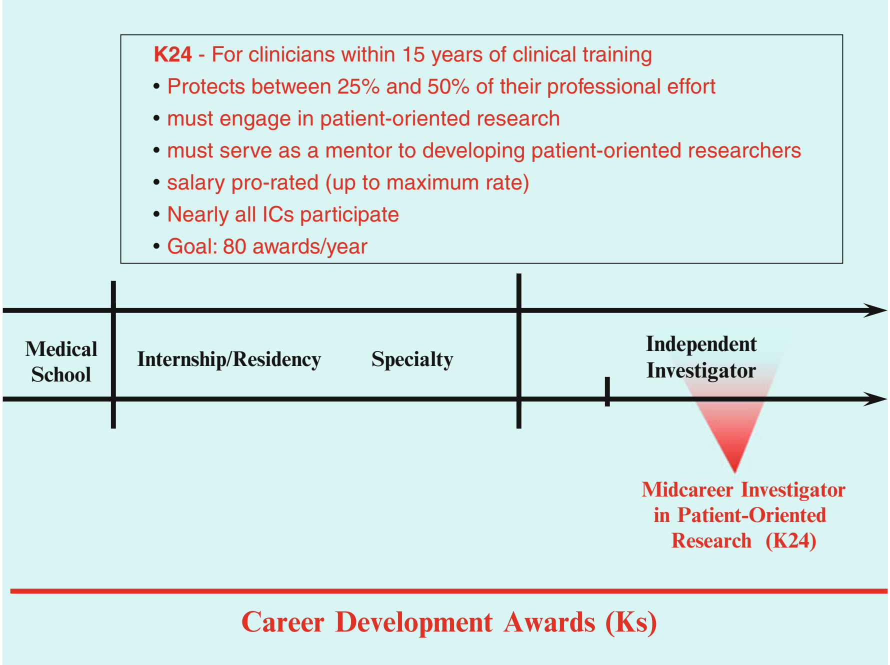
-- For non-patient-oriented research

K23 - Like K08 but for patient-oriented research

– Goal for K23: at least 80 awards/year

Medical School
Internship/Residency
Specialty
Independent Investigator
Mentored Clinical Scientist Development Award (K08)
Mentored Patient-Oriented Research CDA (K23)
Career Development Awards (Ks)

Fig. 21.2 A description of the K08 and K23 awards



K24 - For clinicians within 15 years of clinical training

- Protects between 25% and 50% of their professional effort
- must engage in patient-oriented research
- must serve as a mentor to developing patient-oriented researchers
- salary pro-rated (up to maximum rate)
- Nearly all ICs participate
- Goal: 80 awards/year

Career Development Awards (Ks)

Fig. 21.3 A description of the K24 awards

The K23 application should be very detailed about the mentor's role and responsibilities, how the mentor's area of expertise relates to the research interests of the applicant, how often the applicant will meet with the mentor (and co-mentors), what will happen during those meetings, and how short-comings in the applicant's performance will be addressed. The mentor, on the other hand, should provide the same information, as well as extol the mentor's virtues with prior mentoring activities.

Typically, career development applications should also contain information about formal coursework that will be taken in support of the applicant's career plan, and ideally one that will lead to a degree, such as a Master of Science Degree in Clinical Research (a K30 supported Program). Ideally, the applicants plan will include both an Internal as well as an External Advisory Committee which is formed to provide an objective review of the candidate's progress. More details are spelled out in the grant description, but I have highlighted key components that have been problematic in K23 grants that I have reviewed.

The K24 is a senior non-mentored award that is a natural extension once the K23 is completed. It allows for funded protected time to mentor junior investigators, particularly those seeking a K23 award.

In summary, a number of pitfalls face the junior faculty member interested in a career in patient oriented research. A good mentor/advisor can be of enormous help in guiding young researchers toward their career goals. Unfortunately, many mentors/advisors, acting as role models have fallen into the same traps that they should be preventing in a new researcher, so the mentors role-modeling is somewhat tarnished. We agree with Grigsey that five of the most important pitfalls in the mentor-mentee relationship are: committing to excessive service time; 'diffusion and confusion' i.e. a new faculty member has no clue as to what is or is not a priority without a good advisor guiding them; lack of mentoring/advising; exploitation by other faculty; and, lack of discipline and perseverance.

References

1. Bozeman B, Feeney MK. Toward a useful theory of mentoring: a conceptual analysis and critique. Adm Soc. 2007;39:719–31.
2. Feldman MD, Arean PA, Marshall SJ, Lovett M, O'Sullivan P. Does mentoring matter: results from a survey of faculty mentees at a large health sciences university. Med Educ Online. 2010;15.
3. Aubrey B, Cohen P. Working wisdom: timeless skills and vanguard strategies for learning organizations. San Francisco: Jossey Bass Publishers; 1995. pp. 23, 44–7, 96–7.
4. Guidelines for Faculty/Student Interactions. Division of Epidemiology Faculty Advising Handbook. http://www.sph.umn.edu/pdf/epi/support/docs/Advising-Manual-10.pdf. Accessed 24 Oct 2013.
5. Binkley PF, Brod HC. Mentorship in an Academic Medical Center. Am J Med. 2013;126:1021–5.

Chapter 22
Presentation Skills: How to Present Research Results

Stephen P. Glasser

Speech is power; Speech is to persuade, to convert, to compel

Ralph Wald Emerson

I know from experience that "sometimes it is better to be quiet and be thought a fool than to open your mouth and remove all doubt"…

Abraham Lincoln

Abstract This book is about designing, implementing and interpreting clinical research. This chapter is aimed at a discussion of how to present the research that has been performed. Although almost no one currently disagrees that a formal curriculum in research methodology is critical for a new investigator, the manner in which the results of a study are presented is presumed to be obvious, and training in the art of presentations is much less common. The belief is that good speakers are born, not made, and this is no more true than good researchers are born and not made. And so, the methodology of presentations should be an important part of a young investigators training. This chapter provides an introduction to delivering an effective presentation.

Keywords Presentation structure • Stages of a speaker • Presentation audiovisuals • Question and answer period

This book is about designing, implementing and interpreting clinical research. This chapter is aimed at a discussion of how to present the research that has been

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_22,

Table 22.1 What does an audience remember?

In a lecture given by a brilliant scholar with an outstanding topic and a highly competent audience,
–10 % of the audience displayed signs of inattention within 15 min
–After 18 min 1/3rd were fidgeting
–At 35 min everyone was inattentive
–At 45 min trance was more noticeable, some were asleep and a few were reading a newspaper

performed. Although almost no one currently disagrees that a formal curriculum in research methodology is critical for a new investigator, the manner in which the results of a study are presented is presumed to be obvious, and training in the art of presentations is much less common. The belief is that good speakers are born, not made, and this is no more true than good researchers are born and not made. And so, the methodology of presentations should be an important part of a young investigators training. The ability to communicate effectively is a key to professional success. The investigator who wants to express complex ideas, inform, and educate realizes that effective presentations are an important skill. If you are relatively inexperienced and suffer from stage-fright, relax – you are not alone. Public speaking ranks at the top of the list of peoples fears surpassing even the fear of death. But like any skill, public speaking takes training, experience, persistence, motivation and practice. So what makes a great public speaker? I will attempt to answer that question in this chapter.

In a handbook by Foley and Smilansky [1] the authors quote Frost as follows (Table 22.1), ‘*in a lecture given by a brilliant scholar, with an outstanding topic, and a highly competent audience, ten percent of the audience displayed signs of inattention within fifteen minutes. After eighteen minutes, one third of the audience and ten percent of the platform guests were fidgeting. At 35 minutes everyone was inattentive; at 45 minutes trance was more notable than fidgeting; and at 48 minutes some were asleep and at least one was reading. A casual check twenty-four hours later revealed that the audience recalled only insignificant details, and these were generally wrong.*’ How long should a talk be? ‘A speech, like a bathing suit, should be long enough to cover the subject-but short enough to be interesting’.[1]

[1] The majority of this chapter was taken from personal experience and extensive notes that I had taken from a large number or Presentation Skills Workshops that I have attended. Although I cannot give specific credit for individual pieces of information, I can credit the Instructors of those workshops as follows:

(a) Sue Castorino, President, The Speaking Specialist, Chicago, IL, 1993.
(b) Gerald Kelliher PhD, Associate Dean, Medical College of Pennsylvania.
(c) Eleanor Lopez, Let’s Communicate Better, www.eleanorlopez.com
(d) Power Speaking, and More, Joyce Newman Communications Inc.
(e) Jerry Michaels-Senior Consultant CommCore Communication Strategies.
(f) Science and Medicine Canada, Presentation and Platform Skills Workshop, 1992.
(g) Wyeth Ayerst Laboratories, Ciba-Geigy, Schering, Pfizer, and KOS Pharmaceuticals for sponsoring many of the Presentation Skills Workshops that I attended.

Table 22.2 What are the best ways to transmit information?

	Efficiency	Convenience	Amount of information
Print	High	High	High
Audio CD	High	Moderate	Moderate
Video DVD	High	Moderate	Moderate
Interactive	High	Moderate	Moderate
Lecture	Low	Low	Low

What is the least efficient way of communicating a lot of information, particularly technical information (Table 22.2)? Think about it, and the answer will probably be the oral presentation. Why? for a number of reasons, the most important being that the ear is a limited learning tool. Additionally, the oral lecture is of low efficiency, is associated with low audience recall, and forces the audience to assimilate the information on the speakers schedule, in contrast to a written document or an audio tape or DVD, where a 'student' can review the information at a time when there are no other deadlines that have to be met, or an upcoming appointment for which they do not want to be late etc. Also, the information can be reviewed and re-reviewed at their leisure, important points underlined, and so on. So what is it about the oral presentation that makes it so valuable? Two things: the rapport the speaker can gain with the audience, and the ability of the audience to ask the 'expert' (as one wit defined as someone who lives more than 50 miles away and has slides) questions. In fact, some studies have shown that how a lecture is perceived is 55 % visual, 38 % related to how the speaker sounds, and 7 %, the content. The cliché goes that a famous professor is introduced, and with much fanfare walks to the podium, calls for the lights to be dimmed, and says 'for my first slide….' thereby removing the 55 % visual component needed to gain the necessary rapport that renders the oral presentation so valuable in the first place. If the lights go down, and you can no longer see the speaker, you might as well have an audio tape playing. Standing behind the podium (a protective mechanism) or leaning on it (a message of disinterest), also takes away from the presentation, so when possible it is to be avoided.

The Structure of a Presentation

The old adage for the outline of a talk is the Introduction to the talk – tell them what you are going to tell them; the Body of the talk -Tell them; and, the Conclusion – tell them what you've told them (Table 22.3). Because your audience is most attentive during the introduction and conclusion, those are really the most important parts of the presentation, and of the two probably the introduction is the key in gaining their attentiveness, and the conclusion is most important for the take home messages. Thus, if possible, memorize the conclusion so you do not have to look at the slide, but rather you can look directly at the audience while you make your concluding remarks.

Table 22.3 Anatomy of a lecture

The Lecture
Introduction
–Opening-use an attention getter
–Do not apologize
Body
–Transitions
–Rhetorical questions
Summary
–Know when to stop
–Most important part of the speech
–Briefly restate main points and then STOP

During the introduction you have a free ride for about 2 min and it is during this time, if you use it wisely, that you need to catch the audience's attention. This author likes to use 'hooks' or 'grabbers' during the introductory comments, such as a joke-but be careful in this era of political correctness this can backfire (I have had it happen to me!). Glasbergen has opined about this noting "always start your presentation with a joke, but be careful not to offend anyone! Don't mention religion, politics, race, age, money, technology, men, women, children, plants, animals, food…" (www.glasbergen.com; 2002). One can also use a short video clip relevant to the topic that can engage the audience and demonstrate to them that you have given thought to the presentation. Self-effacing humor (if not overdone) can be useful, a speaker who can laugh at him or herself gains rapport with the audience.

Some examples of humor follow: Groucho Marx's famous quote of 'Before I speak, I have something important to say'; Or, for a presentation about a drug that caused sinus bradycardia, but had no other hemodynamic effect, this author once began a presentation by asking the audience what they thought the most important anti-ischemic mechanism of beta adrenergic blockers was. Most of the audience answered 'sinus bradycardia' after which I responded 'that was my thought as well, but now I am going to tell you about a drug that slows the sinus rate but has no anti-ischemic effects'. Catchy titles for your talk also demonstrate to the audience that you have given some thought to your presentation. Some examples I have used were: 'What do the first flight of the Columbia and quinidine have in common?' (for a talk on re-entry as a mechanism of arrhythmias), or 'What do the Japanese puffer fish and silent ischemia have in common? Alliterations can be catchy also, such as 'Palpitations, Prolapse, and Palpating the Pachyderm' (for this talk on mitral valve prolapse-by the way, I began this talk with the famous poem of the blind man palpating different parts of the pachyderm and coming away with different impressions about what the animal might look like; in order to make the simile of the many ways that mitral valve prolapse can present clinically). Should one consider posing the title of one's talk as a question? Doing so can get the audience thinking, and changes them from taking a passive role in the presentation to taking an active role; thereby gaining more audience attentiveness. Or, one can pose a question in the opening of your introduction such as 'how many of you have patients who have suffered an MI despite the LDL being at goal?' During the author's first exposure to formal training in presentation skills, I was asked to prepare a 5 min presentation. I entitled it 'What do exercise testing and stratigraphy have in common? Digging for answers' – the thesis of the talk being stratification (layers) of risk based on exercise test results, just as a stratigrapher tries to make interpretations based upon the rock layers they observe. In addition to the 'grabber' one should also begin with the thesis of the talk, that is, the 'what's in it for the audience question'. One should also cover the outline of the presentation. The outline should have no more than five points and ideally three points, because studies have shown that after a 10 min presentation, the average listener forgets 25 % of what was said within the first 24 h and 80 % within 4 days [2]. By highlighting the three main points of your presentation and repeating them in the conclusion, you increase the chances that your audience will at least remember the most important points that you wanted to communicate. However, the outline of your presentation should be specific rather than broad. I have heard speakers who have picked up on the point that an outline is important, but unknowingly have 'gotten around it' by using broad general topics. As an example, I heard one speaker, talking on the metabolic syndrome, have an outline that included outline points like: 'I will cover lipid metabolism, the different definitions of metabolic syndrome, and all the treatment options'; when the focus of the talk was really to discuss whether the metabolic syndrome was a precursor to diabetes.

Stages of a Speaker

Almost all speakers have to go though three stages before they become accomplished presenters. The speed with which they traverse these stages depends upon their personalities and whether one follows the precepts outlined in this chapter.

Stage 1 is the fear centered stage. Novice speakers are almost always more nervous than the situation dictates, but being nervous (stage fright) is common to even the most experienced speaker. I remember when Johnny Carson was doing his umpteenth monologue and it was being telemetered as part of the show. Before he went on stage and as he was being introduced his pulse rate surged to 120 bpm! Many novice speakers read from a prepared text to help deal with nerves, but a speech that reads well does not necessarily 'listen' well. The bottom line is that at this stage or any stage, if dressing up like Superman makes you more comfortable, then do it!

Stage 2 is the speaker-centered stage that is characterized by imparting the points you as a speaker wants to make. You have now given enough presentations that there is the appropriate amount of nervousness, you know your subject well, and then you go about presenting everything you know about it. The underlying motivation is probably to impress upon your audience how much you do know, and it is your job to tell them everything! The fact is that for most audiences you will know more about the subject you are presenting then they will (exceptions might be at a national specialty meeting), and this is where another major mistake is made by the stage 2 speaker-assuming a level of knowledge that is really not present and thereby leaving the audience in the dark. This flies in the face of what a good speech should be- clarity, simplicity, and repetition (it is a good idea in talks over 15 or 20 min that after each point you have elaborated in your outline, that you repeat what you just said in one sentence-this entrenches the bullet point that you want them to 'take home'); that is, present a small number of essential ideas, simplicity, and being conversational (see, I just did it) are the attributes of a good presentation. You should strive for keeping your message simple for three reasons: (1) so that you can remember it, (2) so that the audience will understand it, and (3) so that the audience will remember it. Novice speakers and speakers frozen in stage 2 are also notorious for apologizing-apologizing about not having enough time to cover the subject, for not having had time to prepare adequately, for the time of day, month, or season; and, for anything else they can think of. I remember one speaker apologizing for something, then catching himself and apologizing for apologizing! My advice is never apologize! Deal with what you are dealt and go on with it!

Stage 3. It is the third stage that every good speaker should strive for-this is the audience-centered stage characterized by understanding the audience, having a feel for what they really need to know; and, that is dependent upon who the audience is. The fact is that expectations among most audiences, accustomed to the general inadequacy of speakers, are so low that almost any well-intentioned bumbler is, at the very least, accepted – provided that the speaker doesn't drone on too long. With this knowledge, the speaker should now be confident enough in their knowledge of the subject, and relaxed enough that they can control their nervousness. They can now focus on what the specific audience to whom they are presenting absolutely needs to know about the subject-and with almost every subject this can be accomplished with 3–5 main points. It is the integration of the last two stages that makes an excellent speaker, and the approach to message building is fundamental to the art of 'getting to the point'. It is also the stage where you know when to stop! Never, never, never, go over the allotted time, you will not impart any additional information to the audience, and you will antagonize them. I have heard many complaints about talks that have gone on too long, but I have never heard anyone complain about a talk that is too short. One characteristic of the presenter still frozen in stage 2, but knowledgeable enough that he or she knows not to go over time, is to simply take the same amount of material but talk faster; rather than reducing the number of points to be covered. These latter presenter's are sometimes dubbed the 'speed demon' or the 'talking encyclopedia', and this should obviously be avoided.

Audiovisuals

Audiovisuals should be used-but not overused. Most speakers use audiovisuals as a crutch rather than the stepping stones that helps an audience understand the message the speaker is trying to make. Many (most?) speakers also crowd too much information on a slide, and some, knowing that the slides are too crowded, even apologize for it. Comments such as 'I know you cannot see it because the print is too small, but the point I am trying to make is…' If you know it cannot be seen why are you using it for? Epidemiologists are renown for using to much detail in their slides (I can say this because I am one). One of my mentors (Dr. Roy Behnke-referred to as 'Reverend Roy' behind his back because of the way he preached his presentations) used 3–5 slides for an entire Grand Rounds presentation-and those slides had at the most three lines on each. My suggestion is to synthesize the information as is shown in Tables 22.4A and 22.4B. In general, three bullet points per slide is ideal and each slide should have only one unifying idea.

The other common mistake speakers make with slides is related to the use of the pointer. As an experiment one day, watch the eyes of the audience as the speaker uses the pointer like a weapon and is roaming all over the slide instead of holding it steady on the point that they wish to emphasize. As the eyes follow the pointer the listener is distracted from the point that is being made. In fact, if you use a limited number of lines per slide, you can also minimize your use of the pointer, minimize pointer wander, and for those of us who are red-green color blind, it will not matter that one cannot see the red dot from the pointer in the first place. Four types of laser use are (Table 22.5):

- The circle
- The underline
- The back-handed flick
- The epileptic-seizure inducer

Remember,

- DO NOT POINT AT EVERYTHING
  - Not everything is equally important
  - Your voice can provide emphasis too

An accomplished speaker arrives at the venue early enough to become familiar with the AV equipment so that they do not stumble around trying to control the lights (remember to keep the lights as high as possible while ensuring that the slides can be seen by the audience). Reviewing the slide advancement mechanism (hopefully on a PowerPoint or related computer presentation format) is also important so that when their actual presentation begins there is not a lot of stumbling (recall the importance of the opening impression one makes on the audience). As Glasbergen points out, considering "what software would you recommend to give my presentation so much flash and sizzle that nobody notices that I have nothing to say" is not the way to go (www.glasbergen.com; 2002).

The Question and Answer Period

The two main fears about the Q and A are that no questions will be asked, or that questions will be asked for which you do not know the answer. To elicit questions, be invitational such as 'I have been looking forward to your questions' or 'I would

Table 22.4A An example of a table that should not be used in a presentation

	Whites				Blacks			
	Model 1w		Model 2w		Model 1b		Model 2b	
	OR	p‡	OR	p‡	OR	p‡	OR	p‡
Region (Non belt is reference)								
Belt	1.05	(0.93, 1.18)	1.06	(0.95, 1.19)	0.78	(0.65, 0.95)	0.77	(0.65, 0.94)
Buckle	0.89	(0.78, 1.02)	0.91	(0.79, 1.03)	0.91	(0.72, 1.14)	0.90	(0.72, 1.13)
Age (per 10 years)	1.43	(1.35, 1.53)	1.48	(1.40, 1.57)	1.30	(1.18, 1.44)	1.32	(1.21, 1.45)
Male	1.75	(1.56, 1.93)	1.73	(1.56, 1.92)	1.78	(1.49, 2.13)	1.86	(1.56, 2.21)
BMI (per 5 kg/m2) income	1.74	(1.56, 1.93)	1.73	(1.65, 1.85)	1.57	(1.45, 1.71)	1.58	(1.45, 1.71)
(≥75K is reference)								
S2OK	1.26	(1.01, 1.56)			0.95	(0.69, 1.31)		
$20K-$34K	1.24	(1.04, 1.46)			1.05	(0.78, 1.42)		
$35K-74K	1.16	(1.02, 1.33)			1.01	(0.77, 1.33)		
Refused	1.19	(0.98, 1.42)			1.17	(0.82, 1.67)		
Years of education (missing 6)								
(College graduate is reference)								
<High school	1.65	(1.26, 2.15)	1.67	(1.30, 2.15)	1.09	(0.81, 1.46)		
High school	1.22	(1.06, 1.41)	1.27	(1.11, 1.45)	1.05	(0.83, 1.33)		
Some college	1.16	(1.02, 1.32)	1.21	(1.06, 1.36)	1.01	(0.81, 1.26)		
Alcohol Consumption (Non-drinkers are the reference)								
Moderate (0-7 women, 0-14 men)	0.96	(0.86, 1.08)	0.95	(0.85, 1.06)	1.07	(0.88, 1.30)	1.08	(0.91, 1.30)
Heavy (7+ women 14+ men)	1.35	(1.06, 1.72)	1.32	(1.04, 1.66)	2.20	(1.25, 3.80)	2.27	(1.32, 3.94)
Exercise (4+ times per week is reference)								
None	0.94	(0.82, 1.07)			0.74	(0.60, 0.92)	0.74	(0.61, 0.92)
1–3 times	1.04	(0.93, 1.17)			0.88	(0.72, 1.07)	0.87	(0.72, 1.06)

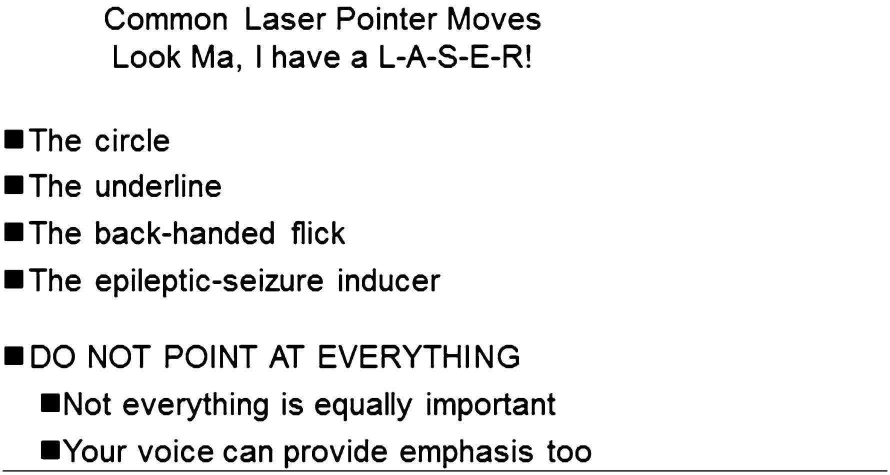
Table 22.4B Compared to Table 22.4A summarizing the information in a more readable fashion is advised

Is Pulse Pressure an Independent Risk Factor for Incident Acute Coronary Heart Disease?

Table: Risks of CHD events associated with pulse pressure levels in participants

Any acute CHD event	**PP <45 mmHg n = 8,099**	**PP 45–54.9 mmHg n = 7,539**	**PP 55–64.9 mmHg n = 4,421**	**PP ≥65.0 mmHg n = 2,850**
Events (n)	139	173	166	203
Unadjusted	1 (ref)	**1.28(1.02, 1.59)**	**2.06(1.65, 2.58)**	**3.82(3.08, 4.73)**
Fully adjusted + SBP	1 (ref)	0.95(0.75, 1.21)	1.15(0.88, 1.50)	**1.56(1.12, 2.18)**

Table 22.5 Common mistakes made when using a laser pointer



Common Laser Pointer Moves
Look Ma, I have a L-A-S-E-R!

- The circle
- The underline
- The back-handed flick
- The epileptic-seizure inducer

- DO NOT POINT AT EVERYTHING
  - Not everything is equally important
  - Your voice can provide emphasis too

be happy to answer any questions'. If there are none, try jumping in with something like 'I am almost always asked about…', and this frequently gets the Q and A going. When a question is asked, keep the answer brief (this is not the time for a mini-talk); and, if you do not know the answer, it is fine to say something like 'I do not know-do you have experience in this area?'- no-one expects you to know everything even if you are 'the expert'. Also, ALWAYS repeat the question so members of the audience who did not hear it are not left out. You can also sometimes rephrase the question so that it is clearer. If the question has nothing to do with the presentation, one can either very briefly address it and then segue into the points you feel are important, or say you would be happy to answer it individually after the Q and A period.

There are a number of other things a speaker can learn about presentations, such as how to answer questions, how to deal with an audience member who is carrying on a conversation during the presentation, the heckler, the know-it-all, the media etc. One should take advantage of courses, seminars etc. that teach these skills. As an example, during a formal seminar on presentation skills, our talks were videotaped and then played back. One of my colleagues-an accomplished speaker-(fortunately it was not me-I had plenty of my own affectations) had his finger in his ear during the entire 5 min mock talk. He was totally unaware that he had done that and even questioned whether the tape had been altered.

As a researcher, it is becoming more and more common to interact with the media about research that you have done (see Chap. 20). Answers to the media have to be even more carefully thought out, because journalists are not only interested in getting the information correctly, but want the ‘headline grabber’ to get people to read about it. They also unknowingly (sometimes knowingly) take things out of context. Despite my experience, I cannot think of an instance where what I intended to be the message of the interview actually came out to my total satisfaction (you might want to think about this when reading an article of someone else who has been interviewed and ‘quoted’). Almost never will a reporter allow you to review beforehand what they are going to print (or edit, if it is a television interview) because they feel they want to maintain their autonomy (by the way, in my view this is more important to them than getting it right). Also, there is a famous (among the presentation skills people) clip from the Bob Newhart show (the one in which he portrayed a psychologist). When he was about to be interviewed before airtime, the reporter was as sweet as sugar, telling him how wonderful his reputation was, what a great field psychology was etc. Then the lights came on, and the interviewer’s first question went something like, ‘Since your field never cures anyone, how can you justify the outrageous fees you charge?’- and it went downhill from there. Hopefully, if you have watched that series, you can imagine how the bumbling Newhart responded.

Conclusion (Table 22.6)

I have found the following points to be critical for a good presentation.

1. A speech that reads well does not necessarily listen well
2. A good speech consists of a surprisingly small number of ideas- do not saturate the audience
3. A secret of an effective speech is simplicity, another is the use of conversational language
4. Content alone will not insure a successful talk
5. Do not apologize about the topic, time etc.
6. Vary the volume of your voice, rate of speaking, etc.
7. Use pauses and inflection along with body movement to emphasize key points
8. Do not exceed your time limit
9. Stand up, speak up, and then shut up
10. Always repeat the question asked, and answer the question briefly.
11. Like your presentation, keep audiovisuals simple with a limited number of points on each slide
12. Keep the room lighting as bright as possible

Griswold outlined 9 ways to “sound like you know what you are talking about” and the list includes: record yourself and play it back, identify and break bad habits

Table 22.6 Twelve commandments of a presentation

A speech that reads well does not necessarily listen well
A good speech consists of a surprisingly small number of ideas
A secret of effective speech is simplicity and the use of conversational language
Content alone will not insure a successful talk
Do not apologize about the topic, time etc.
Vary the volume of your voice, rate of speaking, etc.
Use pauses and inflection along with body movement to emphasize key points
Do not exceed your time limit
Stand up, speak up, and then shut up
Always repeat the question asked, and answer the question briefly
Like your presentation, keep audiovisuals simple with a limited number of points on each slide
Keep the room lighting as bright as possible

(the ums, ers, and uhs), be aware of body language, find your optimal pitch (use your natural speaking voice), speak at the "rate of no mistakes", take advantage of pauses (rather than being concerned about them-pauses are one of the tricks used by great speakers) focus on the continuity of a phrase (think about speaking from one punctuation mark to the next), remember to breathe, and let your enthusiasm show [3]. Additional cautions include speaking to quickly, speaking to quietly, trailing off at the end of phrases, and speaking in monotone

References

1. Foley RP, Smilansky J. Teaching techniques, a handbook for health professionals. New York: McGraw-Hill Book; 1980. p. 1–14.
2. Garson A, Gutgesell H, Pinsky WW, McNamara DG. The 10-minute talk, slides, writing, and delivery. Am Heart J. 1985;111:193–203.
3. Griswold A. 9 ways to sound like you know what you are talking about Business Insider. www.businessinsider.com/common-speaking-mistakes-2013

Index

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4,