Statistical Monitoring of Clinical Trials

Lemuel A. Moyé

Statistical Monitoring of Clinical Trials

Fundamentals for Investigators

With 69 Figures

Springer

Lemuel A. Moyé
University of Texas
Health Science Center at Houston
School of Public Health
Houston, TX 77030
USA
moyelaptop@msn.com

Library of Congress Control Number: 2005931408

ISBN-10: 0-387-27781-1 e-ISBN 0-387-27782-X
ISBN-13: 978-0387-277813

Printed on acid-free paper.

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springeronline.com

To Dixie and the DELTs

Preface

A preface is an opportunity for you and me to share an amiable conversation before the serious work starts. If you give me a moment, I will share with you my motivations for writing an introductory text about the statistical monitoring of clinical trials, a staple of modern research efforts in heatlhcare.

I am pleased to have been involved in clinical research for eighteen years. Many of my efforts focused on preparations for and presentations to Data Monitoring Committees (DMCs), each of which was tasked with overseeing the conduct of a particular clinical study. During these activities, I have spoken with many clinicians about the epidemiology and biostatistical foundation of this mode of clinical research.

In my experience, nothing confuses a DMC member as do these so-called "stopping rules" for monitoring the conduct of a healthcare research study. The idea of prematurely ending a study makes intuitive sense to the clinical members of the committee. The rules themselves with their arcane terminology are the problem. Descriptions of "group sequential procedures" and "stochastic curtailment" provide no useful handholds for the clinician working to understand this slippery but essential subject. The fact that neither medical school nor residency curricula discuss any of the details of these procedures is one possible explanation for the continued lack of understanding among clinicians. In general, the non-statistical members of newly conceived DMCs in the 21st century are just as confused about statistical monitoring guidelines as were their clinical predecessors who sat on DMCs in the 1980s.

A major reason for this continued confusion is that clinical investigators, although blessed with the motivation to do research, commonly do not have strong mathematical backgrounds. Although many have worked hard to develop the basic understanding of epidemiology and biostatistics necessary to be an effective investigator, the underlying mathematical details of commonly used monitoring procedures as frequently presented remain beyond the scope of their training.

Of course, the statistical literature has much to say on the subject of monitoring rules in clinical research. Beginning with the manuscripts of Armitage and Wald in the 1940s, the statistical treatment of this topic slowly expanded until the late 1970s, when it exploded. The recognition of the importance of the monitoring of clinical research, in concert with the complexity of the underlying mathematics has attracted the best and the brightest of biostatisticians. Their devotion to the study of the underlying mathematical structure of monitoring procedures has resulted in a body of knowledge that is both evolutionary and illuminating. However, because it tends to be scripted in the technical and exclusionary language of advanced mathematics, the writing tends to enlighten only the sophisticated analyst.

The text by Christopher Jennison and Bruce Turnbull [1] is a fine example of a comprehensive treatment of a difficult statistical subject.

The technical writing style that has been implemented in the field of "interim monitoring" should come as no surprise. However, work in this area, propelled forward by the strong rowing of capable statistical theorists, can leave the clinical investigator behind in its wake. Required to apply complex processes that they do not understand, the clinical investigator commonly finds little introductory material available. In addition, clinical researchers with no quantitative background have difficulty communicating with biostatisticians or experienced trial methodologists who have much experience but little time to explain these issues to their inexperienced colleagues. Thus, investigators who wish to learn about these mathematical procedures are hard pressed to identify readily understandable source material.

The purpose of this text is to fill that gap. If you know nothing about monitoring guidelines in clinical trials, then this book is for you.

I have chosen to begin this book with a brief history of monitoring rules in clinical research. Although this is the first chapter in this book, it needn't be the first chapter that you read. Being nontechnical, it might be most useful to view its contents as a pleasant oasis in a desert of more complicated discussion. Its considerations of the interactions between scientists serves to convey something about the people who were involved in these important historical efforts. The observation that the epidemiologist Bradford Hill suffered from tuberculosis years before he helped design an early clinical trial to study this disease may be a mere curiosity to some; to others it helps to explain his intellectual fortitude in working with skeptical clinicians.

For the same reason, I have broken up some of the technical arguments that appear in later chapters with an occasional vignette. As my students frequently remind me, it is best to have a joke close by when discussing anything mathematical.

I must confess that this is not a book about the operation of DMCs. That material has been very nicely developed in *Data Monitoring Committees in Clinical Trials: A Practical Perspective* by Susan Ellenberg, Thomas Fleming, and David DeMets (John Wiley & Sons, Ltd., West Sussex, 2002). Their text is very broad in scope, focusing on the DMCs evolution and contemporary operation. Our focus here is on statistical monitoring procedures that these DMCs devise and utilize, not on the DMCs themselves.

One final note. An important segment of the current clinical investigator population is comprised of women. Therefore, I have alternated the use of gender in the hypothetical illustrations offered by this text. Although this is the most illustrative and the least exclusionary approach, it does require mental alacrity on your part as the genders change from example to example.

Lemuel A. Moyé
The University of Texas
School of Public Health
Houston, Texas
July, 2005

References

1. Jennison C, Turnbull BW. (2000). *Group Sequential Methods with Applications to Clinical Trials*. New York. Chapman & Hall/CRC.

Acknowledgments

My understanding of monitoring procedures was not generated by reading complicated treatises, but instead was forged in vibrant and sometimes fiery discussions during Data Monitoring Committee meetings. The distinguished members serving on these panels deserve much of the credit for the educational material this book contains. Special thanks goes to the members of the University of Texas School of Public Health Coordinating Center for Clinical Trials, and to its senior members both past and present. Mort Hawkins ScD, Barry Davis, MD, PhD, and Robert Hardy, PhD. served as the guide rails for me, keeping me on the right track when I tended to veer too far in one direction or the other.

I also acknowledge the members of the Houston SPOTRIAS team. In this effort, we investigators have faced numerous challenges in working to devise useful monitoring guidelines for small pilot research projects, and in doing so, have created through debate and discussion many good ideas. Special thanks to Dr. James Grotta, Andre Alexandrov and Louis Morgenstern for their roles in these conversations.

John Kimmel, the reviewing editor, and several chapter reviewers have provided good questions, challenges, suggestions, and additional references. Their help has improved this book's structure.

Nidhi Rohatgi, my graduate teaching assistant, provided invaluable assistance in clarifying the text and in copyediting. Nidhi was especially helpful in pointing out the need for me to untangle what had become relatively impenetrable tracts of some chapters. Dr. Claudia Pedroza provided fine support for the discussion of Bayesian monitoring guidelines.

Finally, my dearest thanks go to Dixie, my wife, on whose personality, character, love, and common sense I have come to rely, and to Flora and Bella Ardon, whose continued emotional and spiritual growth reveals anew to me each day that, through God, all things are possible.

Lemuel A. Moyé, MD, PhD
University of Texas
School of Public Health
Houston, Texas, USA
July, 2005

Contents

Introduction

Statistical monitoring procedures are the body of computations that aid clinical investigators in determining if a research program should be suspended prematurely. Specifically, these guidelines are used to guide the complex decision to end a clinical study if the investigation is very likely to produce either (1) an early positive benefit, (2) an early indication of harm, or (3) a neutral effect at the time the study is scheduled to end (expressed as stopping for "futility"). Research scientists and members of clinical trial oversight committees rely upon these procedures, colloquially expressed as "stopping rules", but more correcting described as "monitoring guidelines".

Although clinical investigators accept the application of statistical and epidemiologic principles in clinical research, the procedures used to terminate clinical studies often appear opaque to the statistically naïve investigator. Nevertheless, these guidelines have become ubiquitous in healthcare research. In 1998, the Office of the Inspector General of the Department of Health and Human Services mandated that the Food and Drug Administration (FDA) and the National Institutes of Health (NIH) develop such procedures and standards for U.S. trials. In response, the NIH has generated policies to require safety monitoring plans for all phase III NIH-funded studies, and the FDA has issued a draft guidance document on the establishment and operation of the committees that perform such monitoring. In addition, the Institutional Review Boards (IRBs) that govern the ethical conduct of clinical investigation at many research centers developed their own sets of instructions for the application of oversight procedures. These monitoring responsibilities reside in the Data Monitoring Committees (DMCs) of the individual clinical research projects.

This new requisite for formal statistical monitoring of clinical research places clinical investigators in a dilemma. As researchers in a study, they have to satisfy the monitoring requirements of their institutional review board. Alternatively, if they are members of a DMC, then their input into the discussions that calibrate the statistical monitoring device of the study is required. However, these investigators are commonly ill equipped to deal with the issues of modern statistical monitoring of clinical trials. Thus, they are unable to fruitfully engage in the discussion, development, or defense of the use of these tools.

Well-motivated, but statistically unsophisticated clinical investigators can learn the correct use and interpretation of these monitoring procedures when provided with a learning tool that informs them in clear language. This tool would allow them to steadily increase their knowledge of, experience with, and intuition about these procedures. *Statistical Monitoring of Clinical Trials: Fundamentals for Investigators* is this tool. Specifically, it provides the discussion of these statistical devices that clinical investigators need, representing a user-friendly introduction to

monitoring procedures for these scientists. These essential statistical considerations are rarely taught in introductory biostatistics or medical statistics classes.

Chapter One of *Statistical Monitoring of Clinical Trials: Fundamentals for Investigators* provides an overview of the evolution of monitoring procedures in clinical research. Randomized, blinded controlled clinical trials, available for only sixty years, are a relatively new tool in clinical investigation, and remain controversial. The ethical concerns raised by this investigational methodology have called for the interim monitoring of these studies. This demand in turn has generated a relatively new application for Brownian motion, one completely unforeseen by its progenitors, including Albert Einstein.

Chapter Two provides a review of the basic statistical thought process required in clinical research and directly applicable to interim monitoring. The set of circumstances that permit one to generalize the results from a single small sample to a population of thousands or millions of subjects has direct bearing on the successful application of statistical monitoring of clinical trials. These situations and their limitations are discussed in detail. In addition, the foundation principles of statistical hypothesis testing, confidence intervals, and the Bayes approach are each described.

Chapter Three develops the elementary principles of probability that are required to understand the principles behind the interim review of clinical research results. The differences between subjective and objective probability are discussed, and the roles of each in the statistical monitoring of clinical trials are explained. In addition, the concept of probability as an area under a curve is illuminated, with special emphasis given to the normal distribution. Finally, elementary examples of the use of probability for the early termination of a clinical research effort are provided. Chapters Two and Three provide the foundation for the rest of the text.

Chapter Four addresses the need for monitoring procedures in clinical research. This chapter lays out for the clinical scientist the problems that arise when one attempts to use traditional hypothesis testing procedures to draw conclusions about a clinical study's interim results. It provides, through the use of discussion and examples, the elaboration the clinical scientist needs in order to develop insight into the basic behavior of statistical monitoring tools. Investigators have become familiar with the idea of a test statistic's location (i.e., whether the test statistic is greater than 1.96). In this chapter, that notion is supplemented with the observation that a test statistic follows a particular path to arrive at its current location. An examination of that path's properties reveals new information that can provide accurate predictions of the test statistic's location in the future. This concept is new to most clinical investigators, and is elaborated in detail without heavy reliance on mathematics. It is here that the link between Brownian motion and clinical monitoring procedures is motivated.

Capitalizing on the insight provided in Chapter Four, Chapter Five introduces the basic group sequential approach of Pocock and O'Brien–Fleming, followed by discussions of the Haybittle–Peto and Lan–DeMets derivatives. The triangular designs popularized by Whitehead are briefly discussed. Chapter Six develops conditional power in a way that illuminates the circumstances in which a clinical trial may be stopped early for a beneficial finding based on a "look forward" approach.

Chapter Seven describes the use of monitoring procedures to identify harmful effects of the tested intervention. This is a natural introduction to the current use of asymmetric monitoring procedures. In addition, the problem of deciding to discontinue a study because of an unanticipated finding in one of several safety measures is developed. The many unexpected safety considerations that can arise during the study's execution amplify the importance of this issue. This chapter also introduces the notion of stopping a clinical trial early due to "futility".

Chapter Eight provides an introduction to the use of monitoring procedures using the Bayes paradigm. Each chapter ends with a relevant problem set.

This book can serve as a reference text for clinical scientists at all levels of training, being especially useful for healthcare graduate students and junior physician-scientists. Its readers require basic college algebra, plus one course in healthcare statistics. Its contents are of interest to students attending medical schools, graduate schools with an emphasis in healthcare research, and schools of public health. In addition, the contents of *Statistical Monitoring of Clinical Trials: Fundamentals for Investigators* are applicable to workers in health departments, private institutes, and government regulatory agencies. This book is also useful for judges who, not uncommonly, have to learn about the ethical conduct (and, therefore, the ethical monitoring) of clinical research efforts.

This text's incorporation of background material as well as in-depth discussion requires some guidance for its optimal use. There are several sections in Chapters 5, 6, and 7 which have a "*" in their title, signifying that the material is more challenging for students with a weak background in probability. In addition, the appendices, providing some in-depth mathematical development, can also appear formidable to a student with one background course in statistics.

Therefore, this book may be successfully used as the basis for a basic, introductory course on monitoring rules in clinical trials by focusing on Chapters 1 through Chapter 7, ignoring (1) all of the starred sections in Chapters 5, 6, and 7 and (2) the contents of Appendices A through D. However, those with a stronger mathematics background, after reviewing the historical introduction, can move directly to Chapter 4 and proceed through Chapter 8, covering the details of Appendices A through E as needed.

One caveat. Healthcare researchers regardless of their level of mathematical sophistication, should spend some time in Chapter Two, which discusses the statistical reasoning process in medicine. The experience of the author is that, without this review, many researchers unfortunately use statistical monitoring procedures as a tool to identify the "smallest p-value the quickest way" leading to important setbacks in the development of both research programs and research careers.

1

Here, There be dragons….

What will clinical research look like in the year 2065?

The veil of uncertainty shielding our view of the future blocks any detailed response to this provocative question. We might attempt the answer that "in 2065, research will strike the right balance between compassion on the one hand, and the needs of investigational science on the other, their interaction being governed by a overarching ethic." However, this is more of a hope than an observation. Try as we might, we cannot reliably comment on the methodology to be implemented in the mid-21st century.

Just as, we have only the dimmest view of clinical investigation 60 years from now, early clinical trialists working in the 1940s could not imagine what clinical investigation would look like at the end of the 20th century. In the years following World War II, clinical trials fought for acceptance and respectability, struggling to take root in a soil often poisoned by cultural resistance. Many researchers in the 1940s hoped that the "clinical trial" would die a quick death, rubbed out by the ethical dilemmas raised by its use of randomization and treatment blinding.

At that time, linking the random movement of a pollen grain to observations of a clinical trial's treatment effect would have been dismissed as fanciful science fiction. The ideas of Brownian motion were too abstract to be helpful; they were too far removed from any recognizable structure on the clinical research map. These mysterious mathematical tools, like the unknown reaches of the earth located far from Europe on an ancient map, would have simply been stamped with the admonition, "Here, there be dragons."

The following preliminary discussion will etch out the brief history of clinical trials and Brownian motion as these separate fields drifted toward each other. We will see that the mixture of these diverse disciplines has been predictably unpredictable, an observation that we must keep in mind as we plan the trials of the 21st century.

1.1 Clinical Investigation Before the 1940s

Clinical investigation has been a human endeavor for over two thousand years. The most common building block in the edifice of health study is the case report. A case report is a summary of a single patient's findings and the communication of those findings to the medical community. A case series is a collection of case reports, linked together by a common thread (e.g., all of the patients were seen by the same doctor, or each of the patients was exposed to the same agent, e.g., quinine).

It is easy to understand how the growth of general medical knowledge has been propelled by the use of case reports. The delivery of healthcare has been governed by the interaction between a single, concerned, responsible provider and his patient. This relationship is private and privileged. However, it has historically been conducted in isolation, by physicians and nurses widely separated from each other. The idea of a community was well established. However, the concept of a medical community (i.e., a collection of practitioners who worked together to jointly expand their knowledge base) was one that took many generations to develop.

Therefore, medical care was delivered for hundreds of years by practitioners, who, working alone with incomplete knowledge, made decisions that directly affected the lives of their patients, and indirectly, their patients' families and communities. The one, natural learning tool these physicians could use was the active sharing of their experiences among themselves. This served to expand their expertise, suggest alternative approaches to healthcare, and extend their knowledge. This shared experience is at the heart of the case report.

The core thesis of this approach was best captured by Celsus (circa A.D. 25) [1], who stated that "Careful men noted what generally answered the better, and then began the same for their patients." For the next 1900 years, advances in clinical medicine occurred through the combined use of careful observations, clear recorded descriptions, and deductive reasoning. The discovery that gunshot wounds could be healed without the application of burning hot oil [2] demonstrated that a case report-style observation could uncover new information and overturn prior, erroneous principles in medicine. When medical journals began to appear, the primary medical information that they dispersed was that of the case report.[*] Those physicians who had more exposure and experience with a medical issue compiled their case reports together into a case series that they would publish. This continues to this day. Examples are diet drugs and heart valve disease [3] and radiation poisoning [4].

However, case reports have well-established difficulties. Although they reflect very clear and honest observations, the degree to which a single case report represents a general phenomenon in the population can be subject to debate. Even though they are useful, the variability of observations across patients makes it difficult to assess whether one patient's findings summarized in a case report can be easily translated to others.

However, what the case report and essentially all investigative mechanisms in medicine hope to illuminate, by examining both the environment (e.g.,

[*] One of my favorites is an 1822 issue of *Lancet*, whose feature article was titled, "The biggest hernia that I have ever seen in a shipyard worker".

exposure to a toxin or a potential cure) and the patient's response, is the true nature of the exposure–outcome relationship. This true nature could be simply an association, or it could be causal.

An association is the coincidental occurrence of an exposure and an outcome. Its recognition (e.g., the relationship between coffee drinking and pancreatic cancer) typically does not require direct action by the medical community. A causal relationship, on the other hand, signifies that the exposure excites the production of the outcome. This more powerful, directed relationship incites the medical and regulatory communities to action. For example, the conclusion that exposure to citrus fruits reversed the symptoms of scurvy incited action by the British navy to mandate the storage of fresh fruit in the provisions of its crews for long sea voyages [5]. On the other hand, links between the use of cutting and bleedings and the remission of yellow fever were merely associative. Thus, when we as physicians examine a case report's details, we sift through the provided clinical descriptions in order to discern if the relationship between the exposure and the outcome is either causative or associative.

Epidemiologists are specialists who identify the determinants or causes of disease. They have developed criteria that would be useful in ascertaining whether an exposure causes (i.e., excites the production of) the disease. Elaborated by Sir Austin Bradford Hill [6], these tenets are based on a common sense approach to determining causality and are remarkably free from complicated mathematical arguments. These criteria acknowledge that more disease cases in the presence of the risk factor than in its absence raise a causal suspicion. In addition, determining that greater exposure (either by dose or duration) to the risk factor produces a greater extent of disease amplifies our sense that the exposure is controlling the disease's occurrence and/or severity. These two features are important characteristics of a cause–effect relationship.

Other questions posed by Hill permit us to explore the "believability" of the relationship. Is there a discernible mechanism by which the risk factor produces the disease? Have other researchers also shown this relationship? Are there other examples that help us to understand the current exposure–disease relationship? The nine precise Bradford Hill criteria are: (1) strength of association, (2) temporality, (3) dose-response relationship, (4) biologic plausibility, (5) consistency, (6) coherency, (7) specificity, (8) experimentation, (9) analogy. These are well elaborated in the literature [7].

Diligent attempts to determine whether specific case reports and case series can satisfy these causality criteria continue to provide invaluable service to patients and communities. The link between methylmercury exposure and birth defects in communities surrounding Minamata Bay, Japan, [8], and the establishment that thalidomide was the cause of the birth defects phecomelia and achondroplasia [9] are just two 20th century examples of the ability of case reports and case series to establish causal relationships that produced public health action. The identification of (1) the relationship between tick bites and Lyme disease, and (2) the link between new illnesses among postal workers and anthrax exposure in 2001 are recent examples of their continued value.

1.2 Limitations of Case Reports

Although medical knowledge has progressed through the sensitive and intelligent use of case reports and case series, there is no doubt that the illumination provided by these investigational tools is also profoundly limited. There are four major criticisms of the value of case reports and case series in determining the causal nature of an exposure–disease relationship. They are that (1) case reports and case series do not provide quantitative measures of the relationship between an exposure and a disease, (2) case reports do not always rule out other competing causes of disease, (3) case reports are subject to biases of selection (i.e., the manner in which the case report was selected may make it unreasonable to believe that its occurrence reflects an important finding in the population), and (4) measurements made in the case report may be nonstandard. These limitations reduce the contribution of case reports to our understanding of the exposure–disease relationship.

One of the most remarkable deductive failures of case reports was their false identification of the effects of cardiac arrhythmia suppression [10]. In the 1970s, considerable attention was provided to the potential of new therapies (specifically, the drugs encainide, flecainide, and moritzacine) for the treatment of dangerous ventricular arrhythmias. It was believed that these new drugs would be more effective and produce fewer side effects than the traditional, poorly tolerated medications. The effectiveness and safety of these newer drugs were examined in a collection of case series. At first, only the sickest patients were given the new therapy. When these patients survived, the investigational drug was credited with saving the patient's life. However, if the patient died, then the patient was commonly deemed "too sick to be saved" and the drug was not debited for the death.

Based on these observations, despite some opposition, a consensus developed in the cardiology community that patients with arrhythmias would benefit from the use of these new drugs. After a period of intense deliberation, the Federal Food and Drug Administration (FDA) approved the new antiarrhythmic agents. As a consequence of this approval, physicians began to prescribe the drugs not just to patients with severe rhythm disturbances, but also to patients with milder arrhythmias. This new use was consistent with the growing consensus that these drugs would be beneficial in blocking the progression of dysrhythmia from mild heart arrhythmias to more serious rhythm disturbances.

Only after the drugs were approved and on the market was a study carried out that incorporated a control group and the use of randomization. This trial, called CAST (Cardiac Arrhythmia Suppression Trial), demonstrated that, not only did the new therapies not save lives, but their use caused excess mortality [11]. The findings from CAST, demonstrated the lethality of medications whose safety had been "demonstrated" by case series.

1.3 Genesis of the Clinical Trial

By the 1940s, the limitations of the case series as an investigational tool in medicine were evident. However, the evolution of this tool into a device resembling a clinical trial required the patient efforts of the epidemiologist Sir Austin Bradford Hill.

A clinical trial is a medical experiment that is carried out in a unique research setting that must be carefully constructed. The previous section discussed the complicated series of arguments that an investigator must go through in building a causal argument. The clinical trial is the research environment in which many of these properties of the causal argument are already embedded. Upon the beginning of the clinical trial's execution, the only missing feature of the causal argument is the strength of association. This final component is provided by the execution of the study.

Specifically, in a well-designed and well-executed clinical trial, the simple demonstration of a clinically and statistically significant strength of association between the randomly allocated intervention and the prospectively defined primary analyzes is all that is necessary to demonstrate the causal nature of the relationship. This very special situation can only be successfully constructed with (1) a clear statement of the clinical question, (2) a simultaneous focus on epidemiological and biostatistical principles, and (3) disciplined research execution. There are several comprehensive references that discuss in detail the methodology of clinical trials [12,13,14].

The 1930s was a cauldron of new ideas for clinical research. The United Kingdom Medical Research Council's (MRC) Statistical Council and Statistical Research Unit was organized in 1927 [5]. One of its responsibilities was to design and conduct clinical trials in order to investigate promising treatments for modern diseases. The council was adaptive and flexible, opening itself to new and exciting research ideas. One innovative concept was the incorporation of several investigators dispersed throughout a country, all following the same protocol into one research effort. This was the early model for what we now call a multicenter study.

In the mid-1940s, the MRC had the opportunity to evaluate the effect of a new therapy, streptomycin, as a possible treatment for tuberculosis. Streptomycin was a new antibiotic that had not yet demonstrated its effectiveness in clinical experiments. Although it was relatively plentiful in the United States, its availability was limited in impoverished post-war England. The resulting study, conducted by the MRC, was to become the template for the modern clinical trial.

Bradford Hill was asked to design this study. Being both an epidemiologist, as well as a patient who had tuberculosis as a youth,* he held a special appreciation of the complexity of the work required to conclude that streptomycin would be safe and effective for this disease. Hill wished to develop a research paradigm that would produce a clear and unbiased assessment of the effects of the antibiotic. Beginning with the established notion of an experiment in which the researcher has control over the use of an intervention,† Hill successfully argued for three features of the study that were not commonly used in clinical experiments at that time.

* Hill himself had contracted tuberculosis as a young man. He survived a lung abscess, artificial pneumothorax, and a two-year hospitalization twenty-five years before his pivotal streptomycin study.

† An experiment in which the researcher has control of the intervention is different from an observational study, where the investigator has no control of the intervention. An example of an observational study would be that of John Snow's evaluation of the effect of the source of water on the occurrence of cholera, in which the subjects chose their water source.

These were (1) a control group, (2) an external rather than an internal method of selecting the therapy for each individual patient, and (3) blinding, or a procedure to mask both patients and physicians to the identity of the therapy to which any particular patient was assigned [15]. The modern clinical trial emerged from the first attempts to apply these innovations [16].

It is these three features that, in combination, differentiate the clinical trial from other forms of clinical investigation. However, the incorporation of the use of a control group, the random allocation of therapy, and blinding, so essential to the transformation of the clinical experiment into a modern clinical trial, was fraught with controversy. Hill's proposal for their incorporation produced dissension among the clinicians involved in this tuberculosis study. Before we discuss the strong reactions of the research and medical communities to these devices, a reaction that grew to require the need to monitor these studies, we must say a few words about these tools and their intended purposes.

1.4 The Requirement for Control

In the 1940s, the need for a control group was not self-evident to clinical investigators, and it was still common to research potentially new therapies without having patients as comparators. An example was the evaluation of penicillin, in which many of the early studies were conducted without a control group.

There were two main justifications for the absence of control groups in clinical research. The first was the belief that, when the treatment effect was large, then a comparison group would be unnecessary. The second was an ethical one; withholding an experimental treatment was unjustified and harmful when the natural history of the disease (e.g., tuberculosis) was associated with profound morbidity and mortality.*

In this environment, Hill's argument for the inclusion of a control group was not well received by the clinicians who would carry out the study. Those who believed that streptomycin could only have a beneficial effect argued forcefully against the need for a control group. These investigators knew the natural history of tuberculosis; including a comparator group would not substantially add to the body of knowledge concerning the fate of these ill patients. On the other hand, streptomycin's effects were not complete unknowns because the drug had already been partially evaluated in the United States. Why, they asked, withhold a therapy from

* This idea of control group was turned on its head in the Tuskegee syphilis experiment, in which a known, effective therapy was deliberately withheld. For forty years between 1932 and 1972, the U.S. Public Health Service (PHS) conducted an experiment on 399 African-American men in the late stages of syphilis. These men, for the most part illiterate sharecroppers from one of the poorest counties in Alabama, were never told what disease they were suffering from or of its seriousness. Informed that they were being treated for "bad blood," their doctors had no intention of curing them of syphilis. The data for the experiment was to be collected from autopsies of the men, and they were thus deliberately left to degenerate under the ravages of tertiary syphilis—which can include tumors, heart disease, paralysis, blindness, insanity, and death. "As I see it," one of the doctors involved in the study explained, "we have no further interest in these patients until they die." Additional information is available from http://www.infoplease.com/ipa/A0762136.html.

ill patients (likely to die using the standard treatment of care), that was probably safe and could help them?

Hill countered that streptomycin had been incompletely studied to date and must be considered to have unknown effects. If, he argued, the safety and efficacy of streptomycin had already been established, there would be no need to re-evaluate the drug in England.

This scenario was especially disturbing to clinicians, because one of the worst things that they could do would be to give patients with a serious illness a drug that exacerbated their condition. The only way that they could remove the possibility that streptomycin could have harmful effects was by examining patients who would not be exposed to the drug. By helping the investigators to appreciate the limitations of their knowledge about streptomycin therapy, they opened themselves to the idea that streptomycin could be harmful. Investigators discovered that an important new ethical action for them would be to separate their belief about the need for a therapy from their objective knowledge about that therapy's effects. Those who could not would have a difficult time working in the clinical trial era, an observation that is true to this day.

Hill also believed that the high level of efficacy produced by a new therapy could be misleading if that same high level was also seen in the control group. He later demonstrated the importance of a comparison group by revealing that a high success rate for the use of antihistamines to treat the common cold was matched by similar striking findings in a control group [5].

However, acknowledgment of the need for a control group in the tuberculosis study begged the question of which patient should receive the streptomycin as opposed to the control group therapy. As difficult as the fight to include a control group was, the struggle between the clinicians and Hill over therapy allocation would prove to be tougher.

1.5 The Dilemma of Randomization

The random allocation of an experimental intervention is a hallmark of modern experimental design.* The use of random treatment allocations was catapulted to prominence in the mid 1920s by the statistician Ronald Fisher [17,18]. Although Fisher's name is most commonly associated with the use of inference testing in statistics (about which we will have more to say in Chapter Two), he was also one of the pioneers of the use of randomization in research.

Because Fisher worked in agronomy, the first research applications of the random allocation tool were in agriculture. Under Fisher's guidance, new agrarian interventions (e.g., investigational seed formulation or new fertilizer compositions) were allocated randomly to different plots of ground of equal area distributed across the fields. This mix was carefully controlled so that each plot of ground was as

* It is important to distinguish the random allocation of therapy from the random selection of subjects from the population. The random selection of subjects from the population is used in creating the sample, helping to ensure that the sample of patients that is selected for the research is representative of the population from which the sample was selected. The random allocation of therapy occurs after the individual has been selected for the sample. It uses the rules of probability to determine what therapy the patient receives.

likely to receive the new treatment as it was to receive the standard. The resulting patchwork of intervention and control applications helped to ensure that there were no differences between the plots that received the new applications from those that received the standard treatment. Because characteristics of the plots (e.g., proximity to each other, soil moisture and content, insect infestation) did not determine the plot's treatment, these characteristics were removed as possible explicators of the differences in crop yields. This idea of random allocation rapidly took root in agrarian research.

Several years passed before clinical investigators began to explore the possible utility of this procedure for their own work. However, unlike in agrarian research, ethical issues quickly arose in the clinical research arena. It was common for physicians to select the treatment of the research subject. This decision process was simply a natural extension of the habit pattern of physicians in practice who chose the medication for their patients. Therefore, both patients and physicians were comfortable with this historical approach to treatment allocation in clinical research.

Nevertheless, traditional motivations for the therapy allocation contained capricious elements. Inextricably embedded in the decision process were judgments based on the patient's characteristics (e.g., their severity of illness, gender, ethnicity, or financial status). As long as the selection criteria considered characteristics of the patient, it would be impossible to clearly attribute the result seen at the end of the research to the therapy itself.* The random allocation of therapy would solve this problem by creating the environment in which the only difference between patients who receive the intervention and those who did not is the intervention itself, the attribution of effect would be clear [19].

Early efforts at implementing this procedure in clinical research were first attempted in the United States. In 1931, twenty-four individuals who were institutionalized at the Detroit Municipal Tuberculosis Sanatorium were recruited for a study [20]. These cases were individually matched, producing twelve pairs of patients. For each pair, a coin was flipped, and the result of the toss determined which patient of a pair of two would received the active therapy (sanocrysin and sodium-gold thiosulfate injections) versus control group therapy.

Seven years later, 1640 subjects at the University of Minnesota volunteered to receive one of four treatments (three treatments were vaccines, the fourth was a placebo) for the prevention of the common cold. Each student believed that he had received a vaccine when, in fact, the therapy that he received was randomly selected [21]. However, many physicians rebelled against this concept of allowing chance to select the therapy of choice, and the random selection mechanism was prevented from entering the mainstream of clinical research for two decades.

The idea of randomization as a tool re-emerged in Hill's tuberculosis clinical trial fifteen years later. What Hill sought was an allocation mechanism that did not consider personal data, and he believed that the only alternative selection mechanism would be a random one. However, Hill's suggestion that randomization

* This is because those factors that determined therapy allocation would be confused, or confounded with the therapy selection; this confusion makes it difficult to attribute the differences in clinical findings between the control and treatment group to the therapy.

be used in the tuberculosis study ignited a firestorm of debate among its investigators. Physicians could understand the problems generated by poorly planned therapy allocation decisions (e.g., giving the active therapy to only men, and control therapy to only women). However, the notion of making a therapy choice based on the flip of a coin was alien to most, and abhorrent to some.

The motivations for their strong feelings are clear, and resonate to this day. Physicians are trained to be patient oriented. This patient orientation leads us to bring the best of our knowledge, training, experience, and expertise to the patient's bedside. Specifically, when we construct a treatment regimen for a patient, we do it using all of our knowledge about the patient on the one hand, and our expertise with medications. The resultant treatment plan is custom-made for the patient. Woe to the physician who, at the bedside, in front of the patient's family, flips a coin to determine what therapy the patient will receive!

Yet flipping a coin is exactly what randomized therapy is. Hill was obliged to patiently and repeatedly explain to skeptical clinicians what the word "random" really meant. To most clinicians and laymen, then and today, a random process is one that is unplanned, unpredictable, and haphazard. To them, weather could be random, but not a patient's therapy. However, to Hill, random meant a systematic approach in which probability, governed by well-understood laws of chance, would be allowed to prevail. Hill patiently explained that by using chance rather than choice to select the treatment assignment [22], the experiment would provide the independent assessment of a therapy effect, allowing one to "equalize in the two groups the distribution of other characteristics that may be important" [23].

Although Hill had won the fight to include a control group in the tuberculosis study, there is controversy about his success in incorporating the random allocation of therapy. Some suggest that he followed a formal randomization procedure using envelopes completed at a central office that contained each patient's therapy assignment [5]. Others claim that Hill was unsuccessful in persuading the clinicians of the advantages of the random allocation of therapy. These sources argue that the dogged resistance of the physicians to the concept of randomized therapy ultimately led Hill to set the randomized approach aside, replacing it with a strategy of alternating therapy (i.e., the first patient gets active therapy, the second gets control therapy, etc.), a strategy that was more palatable to the investigators [24]. In either case, the trial could only proceed when he avowed to accept a full share of the ethical responsibility for these new trial designs. This willingness on his part was an important reason why clinicians agreed to participate in the studies that Hill designed [5].

Ultimately, the idea of the random allocation of therapy has embedded itself into good clinical trial methodology. However, there continue to be difficulties with its acceptance by some workers, as the following event demonstrates.

> In a randomized, unblinded, multicenter trial designed to compare the effect of different strategies for reducing diastolic blood pressure on the occurrence of strokes, a nurse with established clinical credentials was placed in charge of randomizing patients to either control or active treatment at one clinical cen-

> ter.* One of the patients recruited into this study was an elderly gentleman. Although the patient met the eligibility criteria for the clinical trial, he suffered from several comorbid, cardiovascular conditions. The nurse accepted him into the program, followed the randomization procedure, and entered him into the control group. During the subsequent follow-up visits, the nurse and patient became friends. Shortly thereafter, the patient experienced a clinical endpoint and subsequently died.
>
> The nurse was genuinely saddened by her friend's death, and gave his demise important consideration. She reviewed her previous decision to follow the randomization scheme that had assigned him to receive the control therapy, now wondering whether she was involved in, if not responsible for, his death. After some reflection, she concluded that her patient should have received more aggressive treatment for his hypertension. Deducing that it was the patient's comorbidities, in combination with the absence of active therapy that killed him, she resolved that clinically ill patients would never receive control group therapy at her center. From that point on, any patient who, in her estimation, had not only hypertension, but suffered from other related conditions (e.g., congestive heart failure, diabetes mellitus, or a prior heart attack) would receive active therapy. If the randomization procedure suggested otherwise, then she would merely alter it in this regard.

The outcome of this decision to use active therapy in the sicker patients at this one center was predictable. This allocation of therapy produced a "canceling out" effect, where the beneficial "positive" effect of the medication was canceled by the "negative" effect of the comorbidities' presence. Because this cancellation did not take place in the control group, a systematic bias was now in place that would underestimate the effect of the active antihypertensive treatment. Undoing the randomization process had confused the effects of the therapy with those of the comorbidities, diluting the effect of the medication on the stroke rate at this center.

Although it is easy to criticize this nurse, careful consideration reveals a deeper, more fundamental issue than the mere inappropriate use of her authority. This nurse's only wish was to deliver the best possible care that she could for her patient. However, she was unable to separate her complete belief in the therapy from her true lack of knowledge of the treatment's effects. Comfortable in her belief, this nurse could not stand idly by while a machine made what, in her view, were inappropriate treatment decisions. Her reaction resonates with physicians and nurses who come into research with a strong practice background.†

---

* This occurred before the days of computer-generated randomization procedures, that were instituted for, among many reasons, increasing the difficulty of violating the randomization protocol.

† The problem is much less common among the new generations of clinical trial methodologists, that is, research investigators and their project managers.

Clinical trial methodologists have effectively and persuasively argued that randomization is necessary in clinical trials [25]. In addition, advances in its implementation have been developed (stratified randomization and adaptive randomizations are but two examples) to more flexibly incorporate its advantages into clinical studies. Nevertheless, many of the clinicians whose patients are selected for these studies continue to struggle to understand the necessity of a procedure that appears to be the antithesis of the good practice of medicine. Nevertheless, randomization is the only currently available procedure ensuring independence between a patient's characteristics and their research therapy allocation.

1.6 Blinding

The final adaptation that Bradford Hill introduced into the streptomycin study was a blinding mechanism that masked knowledge of the therapy assignment. In his tuberculosis study, patients were not told what treatment they were receiving. In fact, these patients were not even told that they were participants in a study! [5]* Although this last adaptation is unacceptable in our contemporary research environment, the utility of blinding is uncontested.

Blinding in a clinical study protects the study from influences that can distort the size of the treatment effect. In the previous section, we stated that the motivation for the use of the random allocation of therapy in a clinical trial is to ensure that the only difference between subjects who receive the intervention to be studied and those who do not is the therapy itself. Thus, at the time of the therapy assignment (commonly referred to as the baseline), the distribution of all patient characteristics (e.g., demographics, lifestyle, previous medical history, and physical examination findings) is the same between the two groups; the two groups of patients are equivalent except for the therapy exposure.

Unfortunately, beginning a clinical trial with equivalent patient groups does not guarantee that the trial will end with this equivalence property intact. If the investigators are to be assured that any difference that is seen between the active group and the control group at the end of the trial can be ascribed to the randomly allocated therapy, the two groups of patients must not only have equivalent characteristics at the baseline; the patients must also have equivalent experiences during the study (e.g., equal compliance with the assigned therapy) excepting the effects of the intervention. Ensuring this equivalent post-randomization experience is complicated when the patient and/or the physician knowing the identity of the medication that the patient is taking. Blinding is the collection of procedures that restrict knowledge of the treatment identity. Their implementation increases the likelihood that a patient's post-randomization experience will reflect the effect of the intervention and nothing else.

For example, if a patient knows that she is on placebo therapy, she may believe that her condition is more likely to deteriorate than to improve. This will lead to actions that are motivated, not by the action of the study medicine to which

---

* This was not the first time a clinical study was blinded. In the Detroit Sanatorium study discussed previously, patients were not told which therapy they were placed on (sham subcutaneous injections of distilled water served as the placebo therapy in that experiment).