Current issues in clinical trials: standing on the shoulders of Jerome Cornfield

Susan S. Ellenberg*†

Jerome Cornfield was one of the leading biostatisticians of the mid-20th century and made major contributions to the methods and practice of statistics in many areas. One of Cornfield's major areas of interest was clinical trials. He considered and wrote about many of the issues that continue to concern clinical trialists today. His work ranged from philosophical treatises about the approaches to inference from clinical trials, all the way to assessing the details of the conduct of a particular trial to determine how to best interpret the results. It is interesting to see how many of today's 'hot topics' in clinical trials methodology were addressed in Cornfield's works in the 1960s and 1970s. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Jerome Cornfield; clinical trials; Bayesian methods; cluster randomized trials; interim monitoring; adaptive designs

1. Bayesian methods for clinical trial design and analysis [2–6]

Many statisticians, including Cornfield, were inclined toward the Bayesian approach to the design and analysis of experiments, but until the explosion of computing capability in the latter part of the 20th century, making the required calculations was difficult. As we move ever farther into the 21st century, the use of Bayesian methods is becoming more common, particularly in the earlier phases of drug development trials.

A major factor in the adoption of Bayesian methods for clinical trials is the position of regulatory authorities. Traditionally, regulatory authorities have been resistant to the use of Bayesian methods, largely because of the inherent difficulty of setting prior probabilities. A pharmaceutical company that has invested millions of dollars in developing a new compound and has found the early results sufficiently promising to warrant spending millions more on definitive phase 3 trials is clearly optimistic about the
outcome. Regulators, however, have seen many such cases end in negative results, so they are much less optimistic that any given phase 3 trial will yield positive findings. Spiegelhalter et al. [7] suggested that for regulatory purposes, a ‘skeptical prior’, consistent with the regulatory perspective, be adopted; others have suggested that noninformative priors might be most readily accepted by both statisticians and clinicians [8].

The use of Bayesian methods in the regulatory arena was given a major boost in the late 1990s when the Center for Devices and Radiological Health at the U.S. Food and Drug Administration (FDA) announced its intent to encourage Bayesian designs for evaluating new medical devices. The Center for Devices and Radiological Health leadership believed that for medical devices, there was more of a rationale to establish a prior probability of success based on experience with previous devices and that the Bayesian approach would likely prove more efficient in such trials [9, 10]. Many medical devices have been approved on the basis of clinical trials using a Bayesian design [11].

Bayesian designs have been used for drugs and biological therapeutics in the earlier stages of development, and there has been substantial discussion of the possibility of Bayesian designs for phase 3 trials and regulatory decision making [12], but as yet, no medical products other than devices have been approved on the basis of trials using Bayesian designs.

Bayesian approaches have been advocated in other clinical trials contexts for which use of prior information is considered reliable and/or would very likely increase trial efficiency. In pediatric trials, for example, Bayesian methods using prior probabilities based on experience in adult trials is appealing, as it appears quite reasonable to expect a therapy that benefits adults to also benefit children [13]. Simon has proposed the use of Bayesian methods for active control noninferiority trials [14, 15], another area where establishing a prior probability on the basis of an earlier series of trials will be credible.

2. Interim monitoring and adaptive designs [16–19]

My initial work with Cornfield involved developing a computer program to implement his Bayesian approach to interim monitoring, using a statistic he termed the ‘relative betting odds’ [16]. Bayesian approaches to monitoring have been given some modest attention since then [20–23], but frequentist methods remain predominant in the monitoring of interim trial data, particularly in phase 3 trials. Currently, however, there is greatly increased interest in Bayesian clinical trial methods, much of which can be attributed to an emerging focus on ‘adaptive designs’, using Bayesian methods to modify ongoing trials in multiple ways according to the emerging data. These Bayesian adaptive designs are the most recent development in a long history of formal approaches to modify the design of ongoing clinical trials.

The earliest approaches to adaptation in clinical trials were motivated primarily by ethical considerations and addressed the question of when to stop a trial. A paper by Bross describing sequential approaches to medical trials [24] appeared in 1952, a scant 3 years after the modern era of clinical trials is generally agreed to have begun [25]. Armitage’s work on sequential clinical trials was contemporaneous [26]; his book, providing methods for monitoring trial data and terminating them as soon as results could be considered definitive, was published only 8 years later [27]. Throughout the 1970s and 1980s, methods were developed that accommodated practical issues, such as the emerging practice of performing interim analyses on a regular schedule; group sequential designs [28, 29] responded to this reality.

Another type of adaptation addressed the issue of the adequacy of the prespecified sample size. Sample size calculations are based on the expected outcome and the variability of that outcome, in the control group; the projected difference one hopes to see; and the type 1 and type 2 error rates deemed acceptable for the study. More often than not, the basis for projecting the expected outcome and its variability is weak; the investigators recognize that these estimates may be far from what will be observed in the new study. To handle this problem without concerns about invalidating statistical tests for differences in outcomes, Wittes and Brittain [30] proposed that early accrual into a randomized trial could be considered an ‘internal pilot study’: an estimate of the aggregated variance of the outcome measure could be made from this initial subject subset, and the planned sample size could be adjusted if necessary. Wittes and Brittain showed that this approach had a very minimal effect on the size of the ultimate statistical test performed at the conclusion of the trial.

Multi-stage designs to permit more rapid progress from early to late stages of clinical trials have been explored as a way to progress more quickly from early stage to definitive trials of new therapies and to reduce the likelihood that ineffective drugs would move to large and resource-intensive phase 3 studies. Designs have been proposed to ensure better decision making about moving to phase 3 [31–33] and to select one or more agents from a multi-arm first stage to continue on to definitive evaluation in a
second stage [34, 35]. Such designs, some of which are ‘seamless designs’ that eliminate the potentially long delays inherent in stopping one study and designing and initiating a follow-up study, are appealing as a way to increase efficiency of drug development. Recent work on ‘multi-arm, multi-stage’ designs [36, 37] revisits these earlier approaches, and these designs are finding greater acceptance than when initially proposed.

In the 1990s, statisticians began to revisit the question of whether one might use interim data comparisons, not just to decide whether to stop early but to decide whether to enlarge a trial. Although this concept raised the specter of invalid inferences based on continuing a trial until a desirable significance level was achieved, a number of designs were formulated that protected the size of the final significance test to be performed [38–42]. The use of Bayesian methods for continually updating the prior probability emerged as the approach of greatest interest, particularly for trials carried out by the pharmaceutical industry, as great efficiencies were claimed by its proponents [43]. Although Bayesian adaptive methodology has not yet found its way into large-scale phase 3 trials intended to provide definitive evidence of the benefits and risks of an intervention, it is widely used in early phase trials with the intent of speeding up the ‘go/no go’ decision making [43, 44].

Another type of adaptation has to do with changing the weights of treatment assignments as the trial progresses, from the 50–50 allocation at the beginning of the trial to an allocation that favors whichever treatment is producing the better results. These designs, beginning with the early ‘play the winner’ design proposed by Zelen [45] and known more generally as ‘response-adaptive’ designs, have been advocated [46, 47] as representing a more ethical approach to randomization in long-term trials but have not as yet been widely used.

A long-standing problem in clinical trial design is the evaluation of a treatment for a chronic disease when the outcome of interest is long term, but treatment may change on the basis of short-term intermediate outcomes. This is not an uncommon situation and has presented major challenges for serious progressive diseases such as HIV/AIDS, cancer, and neurodegenerative diseases. Murphy [48, 49] and others [50, 51] have developed designs referred to as ‘dynamic treatment regimes’ or sequential multi-arm randomized trial (SMART) designs that are structured according to clinical practice of changing treatment as required but incorporating randomization at each point that a new treatment must be selected and will permit valid inferences about treatment effects. Such designs will likely be more relevant to defining optimal strategies for using available therapies than to evaluate investigational drugs.

3. Cluster randomized trials [52]

Cluster or group randomized trials have become accepted as the most practical (in some cases, the only practical) approach to answering certain research questions [53]. For example, evaluating interventions that must be administered in a group setting, such as family counseling, classroom approaches to conveying information, or public health communication programs disseminated via television or local newspaper advertising cannot be carried out by randomization of individuals; the unit of randomization must be a group or community. Cluster randomized trials have also been advocated for some vaccine trials as a way to evaluate both individual and herd immunity [54]. The fundamental methodological issue for cluster randomized trials is the need to account for intracluster correlation; the acknowledgement that responses within clusters will be more homogeneous than responses across clusters and therefore the amount of information provided by a given number of individuals within a cluster, all of whom receive the same treatment, is less informative than the amount of information provided by the same number of individuals who had each been allocated treatment at random. Cornfield wrote about methods for such trials in the last year of his life, and his single paper on this topic [52] continues to be cited today.

The basic methods for designing and analyzing cluster randomized trials were worked out in the 1980s and 1990s [55, 56], but methodological issues continue to arise. The importance of having a good estimate of the intraclass correlation coefficient has led to considerations of how best to obtain such estimates and how sensitive the sample size calculations are to these estimates [57]. The concept of matched pair designs has been proposed for use in cluster randomized trials, with pairs being randomization units rather than individuals [58]. Such an approach can be carried out effectively only in trials with a large number of clusters [59]. Special issues relating to informed consent arise in cluster randomized trials, as pre-randomization informed consent of all participants may be difficult or even impossible in some cases, such as when the entire communities may be the units of randomization. A number of papers...
have appeared discussing the concerns that arise and possible solutions, both practical and philosophical [53, 60]. Evidence of the emergence of cluster randomization as a no-longer obscure tool is the recent appearance of a CONSORT paper specifically focused on reporting practices for cluster randomized trials [61].

4. Analytic issues in clinical trials

4.1. Missing data

Since the earliest days of clinical trials, statisticians (actually, all knowledgeable clinical trialists) have worried about the impact of missing data on treatment comparisons. Missing data arise accidentally, when trial subjects withdraw from studies or become lost to follow-up; or deliberately, when investigators stop following subjects who stopped taking their assigned treatment or violated the study protocol in some other way that was deemed unacceptable. When the potentially substantial bias that could be incurred from ignoring such dropouts was recognized [62], the ‘intention to treat’ approach in which everyone in a randomized trial was to be included in the analysis regardless of their compliance with therapy was established as the preferred analytical approach. But how to apply the intention-to-treat principle when data were simply missing was unclear. Some ad hoc approaches, such as ‘last observation carried forward’, became widely used and were considered by many to be the solution to the problem when study subjects dropped out and were unavailable for late measurement of effect. In the 1980s and 1990s, more sophisticated methods for analyzing data sets when some data were missing were developed that were based on modeling the missing data mechanism as well as the missing data themselves [63–67]. These newer methods involve the concepts of data as ‘missing completely at random’, ‘missing at random’, and ‘nonignorable’ missing [68, 69]. Data that were missing completely at random could simply be ignored without biasing comparisons; data that were missing at random could be validly estimated using available trial data. The concept of multiple imputation [69] was introduced to account for the decreased precision of inference when estimating the missing data.

Because all methods for estimating and imputing missing data are based on unverifiable assumptions, no method has emerged as a consensus choice (although more is understood about the potential problems in simply ignoring missing data or using simple imputation methods such as the last observation carried forward). In 2010, the National Research Council issued a report on handling missing data that was prepared by a committee including many of the most prominent statisticians working in this methodological area [70].

4.2. Multiple comparisons [2, 16, 71]

The multiple comparisons problem has plagued clinical trialists throughout the modern era of clinical trials. Cornfield dismissed the possibility that frequentist approaches could lead to satisfactory solutions [16], but many statistical procedures have been developed and are in reasonably widespread use, particularly in the regulatory setting. Multiple testing over time has been handled to the satisfaction of most trialists by group sequential designs and their enhancements [29, 72, 73], but there is less consensus about the many other types of multiple testing that arise in the clinical trials setting. The desire to simultaneously evaluate a treatment’s effect on multiple outcomes has led to the development of global tests [74, 75] and has pushed many investigators to avoid the inferential problem by defining a composite endpoint (which of course has its own set of complications [76]). Numerous methods have been developed that permit testing of multiple outcomes with inferences made on each outcome with less conservatism than the classic Bonferroni method. These include weighted stepwise procedures [77–79], resampling methods [80], and ‘gatekeeping’ methods, hierarchical testing procedures that allow hypotheses to be ordered and permit continued testing as long as significant results are obtained [81, 82]. Benjamini and Hochberg [83] introduced the idea of controlling the ‘false discovery rate’, rather than ensuring that the false positive rate for each outcome tested is controlled.

The issue of evaluating data separately in subsets is a particularly vexing multiple comparisons problem [84]. Testing of treatment by covariate interaction has been advocated as the essential precursor to evaluating subset effects, with tests of effects in subsets credible only if such interaction is documented, but clearly this cannot be the ultimate solution as the performance of many interaction tests has its own multiplicity issues; further, the power of interaction tests is typically low when trials are sized on the basis of power for main effects. With the current proliferation of data on genetic characteristics, the number of subsets of potential interest has increased dramatically. Simon [85] has developed
conceptually simple designs to assign part of the overall type 1 error to a particular subgroup of special interest, with the remaining error assigned to the rest of the study population.

5. Statistics and drug regulation

The growth in influence and importance of statistical methods in the regulatory sphere has been enormous over the past half century. In the late 1960s, the FDA established a special advisory committee, the Biometrics and Epidemiology Methodology Advisory Committee, to be consulted on difficult issues related to the appropriate interpretation of data. One of the first topics addressed by that committee was a surveillance system for adverse events [86]. Jerome Cornfield was a chair of that committee. In 1977, the FDA, under a mandate to limit the number of its advisory committees, disbanded this committee and placed statisticians on most of its subject matter based advisory committees. At the same time, the FDA was also substantially increasing the number of statisticians at the Agency, and regulatory agencies worldwide were recognizing the critical nature of appropriate design and analytical strategies in assessing the safety and efficacy of new medical products [87]. In 1990, the International Conference on Harmonisation (ICH), a collaboration of regulatory and industry scientist from Europe, Japan, and the USA, was established to optimize quality and efficiency in the increasingly global process of drug development. Statistical scientists have contributed to many ICH initiatives; in 1998, the regulatory agencies of the three ICH regions adopted an ICH guidance document entitled ‘Statistical Principles for Clinical Trials’ [88]. Statisticians working in (or extensively advising) industry and regulatory agencies worldwide are heavily involved in development and assessment of new methodologies pertinent to the evaluation of new medical therapies, such as adaptive designs, handling of missing data, multiple comparisons, targeted designs based on genetic characteristics, noninferiority designs, and many other topics. Three journals, the American Statistical Association’s Statistics in Biopharmaceutical Research as well as the Journal of Biopharmaceutical Statistics and Pharmaceutical Statistics, have recently been launched and focus on statistical problems in regulated medical research. The FDA, which employs far more statisticians than any other regulatory agency, has its own Statistical Association, and FDA statisticians are active in all statistical societies as well as pharmaceutically oriented groups such as the Drug Information Association.

6. Conclusion

It is of course impossible to address, even briefly, all of the diverse advances that have been made in clinical trials methodology since the days when Cornfield was active in research. I have focused on areas that Cornfield addressed in his own work, with the possible exception of missing data, an area of too much current interest and work to ignore. I suspect that Cornfield’s increasing focus on Bayesian and likelihood-based methods would have led him to be more enthusiastic about improving estimation in the presence of missing data than in the complex strategies for controlling type 1 errors. I think he would of course be delighted to see the growth in interest in Bayesian approaches for trials and in adaptive designs and pleased with the work on methods for, and actual implementation of, cluster randomized trials. I am sure that he would also be pleased to see the extent to which statisticians are taking (as he did) important leading roles both in the conduct of major studies and in the development of research standards and policies. His papers covered research philosophy as well as mathematical exposition and remain well worth reading even today. Those of us who had the great fortune of working with him continue to miss his insight, his wisdom, his keen wit, and the infectious enthusiasm he brought to statistical problems in biomedical research.

References


42. Chen YHJ, DeMets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine* 2004; **23**:1023–1038.


