Contemporary Clinical Trials 45 (2015) 302–310

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

What can we do about exploratory analyses in clinical trials?

Lem Moyé

University of Texas School of Public Health, 1200 Herman Pressler, Houston, TX 77030, United States

ARTICLE INFO

Article history:
Received 2 June 2015
Received in revised form 19 August 2015
Accepted 15 September 2015
Available online 25 September 2015

Keywords:
Exploratory analyses
Clinical trials
Bayes Theorem

ABSTRACT

The research community has alternatively embraced then repudiated exploratory analyses since the inception of clinical trials in the middle of the twentieth century. After a series of important but ultimately unreproducible findings, these non-prospectively declared evaluations were relegated to hypothesis generating. Since the majority of evaluations conducted in clinical trials with their rich data sets are exploratory, the absence of their persuasive power adds to the inefficiency of clinical trial analyses in an atmosphere of fiscal frugality.

However, the principle argument against exploratory analyses is not based in statistical theory, but pragmatism and observation. The absence of any theoretical treatment of exploratory analyses postpones the day when their statistical weaknesses might be repaired.

Here, we introduce examination of the characteristics of exploratory analyses from a probabilistic and statistical framework. Setting the obvious logistical concerns aside (i.e., the absence of planning produces poor precision), exploratory analyses do not appear to suffer from estimation theory weaknesses. The problem appears to be a difficulty in what is actually reported as the *p*-value. The use of Bayes Theorem provides *p*-values that are more in line with confirmatory analyses. This development may inaugurate a body of work that would lead to the readmission of exploratory analyses to a position of persuasive power in clinical trials.

1. Introduction

Clinical trials may be too inefficient to survive this era of diminishing financial investment in health care research.

Such a statement was unutterable ten years ago. Yet the utility of this reliable research tool is now being squeezed by the pernicious combination of two forces — one acute, the other chronic.

The first of these forces is a wave of fiscal conservatism. As National Institutes of Health funding for research declines [1] and pressure grows to divert financial support to smaller programs [2,3] nationally funded multicenter clinical trials require new efficiency and return on investment to justify their existence. The situation is exacerbated by the recent debate over whether some sectors of NIH research are overfunded [4,5].

The second force is internal to the clinical trial itself. Clinical trials generate many analyses, yet only a small fraction of them are held out as persuasive and contributory. The research community expects that clinical trial research will be divided into two broad areas of evaluations; 1) prospectively declared analyses and 2) hypothesis generating or exploratory analyses. Prospectively declared evaluations are themselves partitioned into primary analyses (where type I error is conserved) and secondary analyses that are prospectively declared and in many circumstances can be interpreted unambiguously [6].

The intensive effort required to prospectively design endpoint analyses, manage type I error, and precisely measure endpoints during the execution of the study combine to keep the number of prospectively declared endpoints to a small manageable set. Alternatively, exploratory evaluations — requiring no prospective planning — are numerous. However, despite the larger number of exploratory analyses commonly performed by a single clinical trial, it is the smaller collection of prospectively declared evaluations that currently hold the greatest value to the research and regulatory communities. Standards require that published studies report on all prospectively declared endpoints regardless of their findings [7] and more recently, the federal government has mandated reporting requirements for these a priori planned evaluations [8]. However hypothesis-generating analyses, which represent the majority of assessments in clinical trials, have little persuasive power and follow no reporting guidelines. Thus, the reporting custom of clinical trial results permits most of the analyses the study conducts to remain unreported, inducing a profound inefficiency. Diminishing financial resources make this state of affairs less palatable.

Understandable reasons created this state of affairs. This paper will demonstrate how statistical methodology might begin to reduce the

E-mail address: Lemmoye@msn.com.

http://dx.doi.org/10.1016/j.cct.2015.09.012
1551-7144/

barrier between confirmatory (prospectively declared) and exploratory analyses, allowing in some cases exploratory analyses to have new persuasive power and thereby increase the efficiency of the clinical trial.

2. Background

The penetrating contribution of clinical trials to medical and public health knowledge since their inception in the mid twentieth century is unquestioned. Implementation of statistics and epidemiology has not only amplified the incisiveness of this research methodology, but has fueled advancement in each of these quantitative fields. However, the acceptability of exploratory analyses in clinical trials is based on experience — not statistical theory — and has varied.

The interpretative clarity of the first major clinical trial's results galvanized the public health community to first learn and then wield this research implement. The study by Sir Bradford Hill on the impact of streptomycin on tubercular mortality conducted by the Medical Research Council of Great Britain just after World War II [9,10] revealed that the simultaneous presence of 1) a contemporary control group, 2) randomization of treatment assignment, and 3) some degree of blinding clearly delineated effects attributable to the treatment under study. This design, although criticized by many of the participating physicians early in the groundbreaking study [11], was catapulted to new popularity because of its uncontested results.

Simultaneously, the *p*-value developed by RA Fisher in 1925–26 [12,13], despite some initial derision, [14–16] accelerated to prominence in health research interpretation. This was principally due to the confluence of needs of journal editors, federal grant reviewers, and FDA administrators to choose worthy research products from the plethora of post-war research activity [17]. The combination of the clinical trial (with its simplicity of interpretation) on the one hand with the *p*-value (that combined effect size, variability, sample size, and sampling error into one number) on the other created a new and unbeatable investigative combination in health care research. Results from clinical trials that produced *p*-values <0.05 were accepted with little question by the medical community. The concern expressed by epidemiologists for this uber-distillation of a major research endeavor to one number [18–22] was dismissed by investigators who believed that the clinical trial had earned the rare position of dispensing "truth" based on the "$p < 0.05$" metric from any of its analyses. Any result generated from a clinical trial with a small *p*-value was considered generalizable, and while eminent clinical trialists offered monitories about clinical trial mistakes, they expressed no concerns about this reporting tendency [23,24].

The spectacular findings of the Multiple Risk Factor Intervention Trial (MRFIT) [25] alerted the cardiology and public health communities to the dangers inherent in this reductionist approach. Published in 1982, MRFIT set out to demonstrate that reducing risk factors commonly associated with atherosclerotic heart disease (e.g., hypertension, diabetes, obesity, and smoking) decreased the incidence of heart attacks and strokes. At the study's end, the investigators concluded that their interventions had slightly increased rather than dramatically decreased the incidence of the clinical cardiovascular disease. However, in reviewing their entire dataset, they observed that in the subgroup of hypertensive men with heart abnormalities at rest, larger clinical event rates were associated with the use of antihypertensive therapy [25]. The application of the small *p*-value to a result from a clinical trial (whether that result was produced from a prospective analysis or not) convinced them and their colleagues of the veracity of this findings [26], injecting new doubt into public health initiatives for the treatment of hypertension [27]. However, to the consternation of many, this subgroup analysis with its small *p*-value could not be reproduced in other clinical trials. To a research community that at the time expected "truth" from clinical trials, the appearance of this unreproducible finding was disturbing.

There were other surprises. The Vesnarinone in Patients with Heart Failure Trial [28] identified a sizable mortality benefit in a clinical trial to assess the effect of vesnarinone in patients with heart failure. This finding was overturned by a following clinical trial VEST [29] that demonstrated a small and hazardous effect on mortality attributable to vesnarinone instead of a benefit. The mortality benefit of losartan in heart failure patients, discovered in the Evaluation of Losartan in the Elderly Study (ELITE) clinical trial [30] was reversed by the findings of ELITE II [31] that identified no such effect. The Prospective Randomized Amlodipine Survival Evaluation (PRAISE) [32] mortality benefit attributable to amlodipine in a subset of heart failure patients was reversed by the findings of PRAISE-2 [33].

The angst produced by well-executed clinical trials reversing the findings of other small *p*-value driven, well-executed clinical trials was palpable throughout the research community. Investigators who were trained to believe that clinical trial results were the most solid of all research efforts developed a new permeability to the concept that perhaps not all promulgated findings from these studies were equal. A new metric was needed [34] and some workers began to dissect and separate exploratory analyses from prospectively planned evaluations [35].

The explosive emotions generated by the 1995–97 US Carvedilol controversy revealed the potential losses sustained by trial sponsors as a consequence of changing the clinical trial interpretative paradigm. Carvedilol, at the time an approved treatment for hypertension, was studied as a potential therapy for the treatment of heart failure. Stunning results from the US Carvedilol program [36] suggested that the drug produced a substantial mortality benefit. However, when this result was sifted through the metric of prospective versus non-prospectively declared analyses at a public session sponsored by the FDA, different points of view discounting the overwhelming benefit were aired. This emotive and vehement debate spilled into the medical literature [37,38] followed by full length manuscripts addressing the strengths and weaknesses of the clinical trial methodology [39–42], illuminating the trail of difficulties forged by reducing emphasis on non-prospectively declared analyses in clinical trials.

Clearly, not all accepted this new interpretative mantra. In fact, epidemiologists had long pointed to scientific rationale for conducting hypothesis generating results. They showed that the internal consistency of all of a study's analyses should be examined for support of the study's overall findings. Evaluation of underlying mechanisms of action required to support biologic plausibility were critical to the causal argument [43]. Also, many argued that the role of discovery — visualizing a new and promising scientific relationship for the first time — could not be ignored just because the analysis or finding was not prospectively planned. Compound 2254RP, first developed as an antibiotic, had its more important blood sugar lowering potential recognized only when it produced unanticipated seizures in test patients [44].These "exploratory findings" were later confirmed. Madam Curie discovered radiation, exploratory findings that were also confirmed.

Meanwhile, work proceeded to untangle what was once one of the easiest tasks in medical research — clinical trial interpretation. This stream of investigations beat an ever louder rhythm for change. Clinical trialists offered the notion that the *p*-value did not require replacing, but merely needed a new context. The analyses that were prospectively planned might have a useful *p*-value assessment. Other, non-prospectively declared analyses, even though they were derived from clinical trials would be denigrated to second class status. Labeled as "exploratory," their *p*-values would be deemed uninterpretable. This commonly included subgroup analyses, the examination of dose–response relationships, adjusting therapy effects for covariates, and the evaluation of new "endpoints" that were not prospectively declared.

A corollary of this approach was that a clinical trial whose primary endpoints were not statistically significant could not be resuscitated by a positive finding of any exploratory endpoint regardless of how clinically compelling the case for the exploratory endpoint might be. The FDA codified this thought process through a guidance: