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.

© 2015 Elsevier Inc. All rights reserved.

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/© 2015 Elsevier Inc. All rights reserved.

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:

> "For each clinical trial contributing to a marketing application, all important details of its design and conduct and the principal features of its proposed statistical analysis should be clearly specified in a protocol written before the trial begins. The extent to which the procedures in the protocol are followed and the primary analysis is planned a priori will contribute to the degree of confidence in the final results and conclusions of the trial."Guidance for IndustryE9 Statistical Principals for Clinical TrialsFDA September 1998 [45].

Thus, by the turn of the twenty-first century, clinical trial results were divided along two lines. The first partition was between prospectively declared or exploratory analyses, the latter being reduced to "hypothesis generating." The second partition was between prospectively declared evaluations that had their type I error conserved (primary) and those that did not. Primary endpoints alone would determine whether the clinical trial was positive, null, or negative. Essentially, if the multiplicity issue is addressed, the endpoint is primary. All other prospectively declared analyses are conducted "nominally" i.e., there type I error expenditure is capped at 0.05 without a correction for multiplicity. These analyses are considered supportive if conducted properly [46]. (Fig. 1).

The advice to exploratory analysts of clinical trials is that their work, while worth conducting, had no staying power in and of itself. The role of non-prospectively declared evaluations such as complex general linear modeling, time to event analyses with time dependent covariates, biomarker assessment, and simulation was to create questions, stimulate new thinking and drive future experimental design. Thus, the state of the art advice for 2015 clinical trial design is that the protocol should make a clear distinction between the aspects of a trial that will be used for confirmatory proof and the aspects which will provide data for exploratory analysis [47]. The result of this well motivated advice is that while exploratory analyses are voluminous, they are also commonly seen as valueless since "significant" findings often are not reproducible. Thus, although they commonly outnumber primary and secondary analyses and are the most prolific aspect of a clinical trial, yet remain tightly confined by the lessons of clinical trial history.

2.1. Methodology

There was no underlying statistical theory that drove clinical trial exploratory analyses down to the lower tiers of evidence. Instead, practical observations ruled the day as pragmatic inspections by investigators, regulators, and clinical methodologists demonstrated that these analyses could not in general be reproduced. Experience does not disavow the utility of this empirical decision.

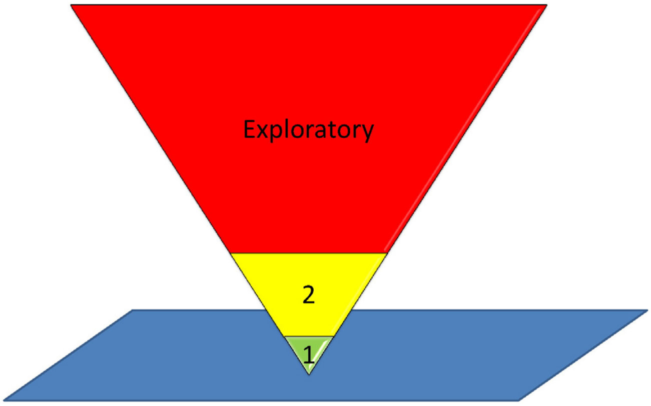

Fig. 1. Hierarchy of analyses in clinical trials. The primary endpoints are smallest in number but the entire weight of the trial depends on them.

However, the absence of a theoretical basis for the relegation of exploratory analyses paints clinical trialists into an uncomfortable corner. With no methodologic basis for the weakness of exploratory analyses, there is no framework on which to construct a new theoretical foundation for their ultimate admissibility into the universe of confirmatory analyses. In the current environment, we have placed flashing hazard lights around exploratory analyses, but have not created the structure that would permit us ultimately to remove the limits from these sometimes illuminating evaluations.

In an attempt to repair this, two major shortcomings of exploratory analyses will be provided. One is nonstatistical, but the second involves a probabilistic decision process that can be parameterized and characterized within the environs of probability theory. The application of this commonly used theory creates a path that when followed, may lead exploratory analyses into the mainstream of clinical trial investigations, thereby increasing the efficiency of the clinical trial.

The principal justification given for the unreliability of an exploratory analysis is the effect of the absence of prospective planning on the precision of the exploratory estimators of effect size. We will call this the logistical rationale. For example, if investigators wish to conduct a clinical trial on the effect of an intervention on heart muscle perfusion, they are obligated to ensure superior quality and high precision images for the endpoint measures, e.g., identifying a core laboratory. These trial design controls reduce endpoint variability and *ceteris paribus* increase power. However, should these same investigators observe at the study's conclusion a treatment attributable benefit for coronary artery disease death, they will be hard pressed to defend the reliability of this unanticipated finding. The absence of its prospective declaration meant that there was no opportunity to organize resources for its reliable estimation. Without prior definition of coronary artery disease death, there could be no *a priori* structure in place for the formal collection of death records and no endpoint committee of specialists to adjudicate findings. In addition, the analysis suffered from an absence of prospective statistical planning that, had it been present would have produced both informative power computations and the minimum number of deaths required to draw a conclusion with statistical regularity.

But, is the logistical rationale the only concern in exploratory analyses? Consider the following thought experiment adapted from Moyé [48].

> An enthusiastic young researcher, Dr. C is interested in demonstrating the progressive deterioration in heart function observed in patients with mild heart failure whose medical management is appropriate. She has designed a research program to examine the changes in left ventricular function over time, with specific attention to examining changes in left ventricular end diastolic volume (LVEDV). As the heart weakens over time, she anticipates that the left ventricle will enlarge and the volume contained therein will increase. Her research design includes a standard sample size computation designed to detect a meaningful increase in LVEDV over the predetermined two year follow-up period.
>
> Upon approval of her study, Dr. C recruits her sample randomly from the population of people with heart failure at her institution, and follows each patient for two years per protocol. Every subject has their heart function measured when they enter the study and again in 24 months. No patients are lost to follow-up.
>
> Her colleagues who have contributed their own patients to Dr. C's investigation are anxious to learn of the conclusions of her study. At the two year time point, Dr. C. examines her data with great anticipation. Comparing the baseline to twenty-four month change in LVEDV for each of her patients, she discovered to her surprise that the anticipated increase in LVEDV did not materialize. However although LVEDV has not increased, there has been a substantial increase in left ventricular end systolic volume (LVESV). She reports

> both LVEDV and LVESV, and claims that her study is positive based on the LVESV finding.
>
> While some of her colleagues applaud her, others voice concern over the lack of a prospective statement for the LVESV analysis. They criticize its prominent location in the paper, and say it should be reduced to exploratory status.
>
> However, Dr. C. replies that she has met her obligation by reporting the prospectively declared and null LVEDV finding. Furthermore, she states that if she had a priori announced that LVESV would be a goal of her study, the study design would have been no different. The state of the art measuring tool for heart volumes (magnetic resonance imaging) was utilized for both endpoints. The variability of the two volumes is the same, and the anticipated change in the effect size of interest is also close between the two measures. In fact, the research effort would have been designed no differently if she has selected LVESV rather than LVEDV. She therefore insists that the "exploratory" LVESV finding be afforded the same status as the prospectively declared LVEDV. Was Dr. C. correct?

This example permits us to set logistics aside to see if there is anything else appreciably different about Dr. C's selection choice for LVESV then there was about the prospective declaration for LVEDV.

Dr. C. chose LVEDV prospectively. She would have (and was obligated to) report on the finding regardless of the size of the test statistic. However, the circumstances are different for LVESV. This finding came to her attention because of the size of the test statistic – had the analysis demonstrated a small value of the test statistic (and a large *p*-value) she is likely not to have reported it at all.[1] The large test statistic that she observed essentially drew her to the finding, answering a question that she did not think to ask prospectively. So, unlike in confirmatory analyses, here it was the test statistic's size and not a prospective decision that determined whether Dr. C. would report the LVESV finding. This is the hallmark of exploratory analyses – and discovery.

Although the magnitude of the LVESV test statistic could have reflected the true state of nature between LVESV and time, there was also a sampling error component. Since the data contain sample-to-sample variability, we know that different samples would provide different "positive" findings just through chance alone. For example, in some datasets, neither LVEDV, nor LVESV would change significantly, but cardiac output would. The random aggregation of the data creates the random findings in the data. And if the variable is selected randomly by the data on which the estimate of effect itself is based then it is fair to ask if this selection process could have an impact on the statistical estimator of effect size. Put another way, is the simple sample mean X no longer optimal if the selection process of the variable X itself was not fixed, but random?

Statistical estimates (e.g., incidence rates, mean differences, relative risks) are based on estimation theory, and their accuracy can be traced to the degree that the assumptions of the statistical estimators remain intact. The derivation of each of these estimators begins with one source of random error that is the selection of the actual data points. But what of the selection process of X itself? Is the estimation theory altered by the random selection of X, and what if X is chosen by the same dataset (and subject to the same effects of random error) as the data points $x_1, \ldots, x_n$ which are the substrate for the estimator? In order to examine this, we can conduct a final thought experiment.

Consider a collection of four random variables, W, X, Y, and Z in a clinical trial. These variables may represent for example blood pressure, level of education, number of hospitalizations a patient has experienced, and death from a heart attack. If we select one randomly and then analyze that variable, then does the random selection of the variable affect its summary estimate?

The answer is (of course) that it depends on the purpose of the analysis. If one is trying to estimate the long term behavior of the process (random selection, and then analyze the variable selected), by estimating the mean and variance of the selected variable, then the answer is "Yes". However, if one is simply trying to estimate the mean and variance of the selected variable, then the answer is "No". Since it is the latter that is conducted in clinical research, the random determination has no effect on the exploratory estimator.

However, this thought experiment is not an exact representation of the exploratory research process. In exploratory analysis, a variable is not first selected and then analyzed. Instead, the universe of exploratory variables in a clinical trial is analyzed, taking each variable one at a time. The random selection takes place not before, but after all variables have been analyzed. What is random is not the decision to analyze, but the choice of which of the exploratory findings to report.

That the decision is random comes as no surprise. Clinical trials commonly make random decisions. For example, the determination that a clinical trial is positive based on a reduction in the prospectively declared primary endpoint of total mortality with a *p*-value of 0.035 is a random determination containing sampling error. It is based on the likelihood of a (random) event that a population in which no mortality benefit occurs produces a sample with the positive data in hand. We compute the probability of this event, determine that this probability is small, and then announce the positive findings. But the value of the test statistic is simply and merely one observation of a random event.[2] Similarly, the determination to end a trial early for efficacy, harm or futility, fueled by Brownian motion concepts [49], is a random decision. So the random decision process is not new to clinical trialists.

What is different in exploratory analyses is that the selection of the variable to be reported is based on the value of the test statistic. For confirmatory analyses, the test statistic reported is based on the *a priori* selection of a variable. To return to our first example, Dr. C had to report the test statistic (and *p*-value) of the prospectively declared endpoint LVEDV because of its prospective choice, but only reported the exploratory variable LVESV because it was significant. The probability that the investigator reports the exploratory variable depends on the value of its test statistic. If the test statistic falls in the critical region, she is more likely to report it. In prospectively declared analyses, the investigator reports the test statistic given the variable. In exploratory analysis one reports the variable given the test statistic.

Thus, "reversing the condition" is one of the concepts that separates exploratory analyses from confirmatory analyses, suggesting that with the right information, we can convert the critical probability in the exploratory realm to a probability closer to the relevant confirmatory *p*-value.

2.2. Parameterization

An attempt will be made here to provide a new estimate of statistical significance within the exploratory paradigm. We will only address the one–tailed testing circumstance, although our results readily generalize

[1] This is setting aside for a moment the notion that journal reviewers may have insisted on seeing both volumes because they each measure heart function.

[2] We use additional epidemiologic criteria (dose response effect, understanding of the underlying mechanism of the result, have other investigators seen this effect) to help assure us that our decision is true and not a chance finding, but at its heart, the decision to declare the study positive is a random one. This may be seen from two perspectives. First, considering one and only one trial, if the test statistics could be tracked moment by moment during the follow-up period of the study, then it would meander over time as a random walk/quasi Brownian motion process. At any one point it is fixed, but the test statistic is random.

In addition, stepping back for a second, if we consider the universe of trials conducted under this protocol, each trial would collect data from different subjects and therefore different life experiences. Thus, the test statistic that is obtained at the end of each of these trials is a random realization of the test statistic that would be produced by the population.

to the two-tailed testing paradigm. This difference will be parameterized as follows.

A major interpretative difficulty introduced by exploratory analysis is that the decision to report an exploratory result is based on the magnitude of that result's p-value. Denote S as the selected variable to be reported, and v as one of the candidate exploratory variables that could be reported. Then $\mathbf{P}[S = v]$ is the probability that the variable v will be reported. Let w be the test statistic produced from the analysis of variable v.

The p-value is simply the probability that the test statistic is at least as large as its observed value w. In the exploratory setting, a relatively large number of variables is examined and the decision to report each of them is based on the value of their individual test statistics. Since the value of the test statistic informs us of whether $S = \nu$, the exploratory scientist is most interested $\mathbf{P}[S = v_i | Z \geq w_i]$, where Z is a standard normal random variable.

The situation is different in confirmatory analyses. In that environment, the p-value does not drive the selection of the variable to be reported. Instead, the variable's analysis is conducted, its p-value computed, and the result $\mathbf{P}[Z \geq w | S = v]$ reported. Since, the result is reported regardless of the magnitude of the test statistic, the investigator simply provide the p-value as $\mathbf{P}[Z \geq w_i | S = v_i] = \mathbf{P}[Z \geq w_i]$ with the conditional statement of the event implied. Thus, a difficulty in exploratory analysis is that while we commonly assume that what one is reporting for the exploratory variable is the classic p-value $\mathbf{P}[Z \geq w_i | S = v_i] = \mathbf{P}[Z \geq w_i]$, what is actually reported is a quantity that confounds 1) the probability that the test statistic falls in the critical region with 2) the likelihood that the report will be made, $\mathbf{P}[S = v_i | Z \geq w_i]$. Our goal here is provide a simple statistical framework that permits one to convert $\mathbf{P}[S = v_i | Z \geq w_i]$ to $\mathbf{P}[Z \geq w_i | S = v_i]$ the latter of which we will call the exploratory p-value or p_e. However, this is only a partial step, since we cannot make the leap to $\mathbf{P}[Z \geq w_i | S = v_i] = \mathbf{P}[Z \geq w_i]$. This independence property of confirmatory analyses is outside the exploratory paradigm. However, the conversion that we will carry out will bring the p-value closer to the realm of that produced by confirmatory analyses. In order to conduct the inversion, we require only simple conditional probability and the invocation of Bayes Theorem.

2.3. Application of Bayes Theorem

We wish to find the probability density function of the test statistic in exploratory analysis. That is from the k pairs of exploratory analyses and endpoints, and their associated test statistics we choose the ith one arbitrarily. If we had in hand $f_w(w | S = v_i)$ the probability density function of the test statistic w_i given the exploratory analysis ν_i was selected, we could compute the exploratory p-value p_e as

$$p_e = \int\limits_{CR(w_i)} f_w(w_i | S = v_i)\, dw_i$$

where $CR(w_i)$ is the critical region for the test statistic w_i. We invoke Bayes Theorem to compute the density function $f_w(w_i | S = v_i)$

$$f_W(w_i | S = v_i) \;\;=\;\; \frac{f_{\nu_i}(\nu_i | w_i) f_W(w_i)}{\int\limits_{\Omega_W} f_{v_i}(\nu_i | w_i) f_W(w_i)\, dw_i}$$

where $f_{\nu_i}(\nu_i | w)$ is the probability density function governing the likelihood that the investigator will report the exploratory analysis ν_i given the test statistic value w, and $f_W(w_i)$ is the probability density of the test statistic (we will assume throughout that w_i follows a standard normal distribution under the null hypothesis).

2.4. The identity of $f_{\nu_i}(\nu_i | w_i)$

This probability density function reflects the inclination of the investigator to report an exploratory result based on the value of its test statistic. It is easy to defend the assumption that the larger the test statistic, the greater the likelihood of reporting and exploratory evaluations whose analyses do not fall into the clinical region have a very low probability of being reported. We begin with modeling this event as.

$$f_{\nu_i}(v_i | w_i) = \frac{\lambda_i e^{\lambda_i w_i}}{\left(e^{\lambda_i b} - e^{\lambda_i a}\right)} 1_{a \, \leq \, w_i \, \leq \, b}.$$

This is an exponential density function normed on (a, b) for $0 \leq a < b < \infty$. We will assume the value a is the lower bound of the critical region (typically 1.96), and the value b its upper bound (a limiting approach will allow us to dispense with b shortly). For example, assume that an investigator has conducted an exploratory analysis on a variable v_1 that produced a test statistic w_1. Then using this model, the probability that an investigator will report that analysis is $\mathbf{P}[S = v_1 | W = w_1] = \int\limits_{w_1}^{b} \frac{\lambda_i e^{\lambda_i w}}{\left(e^{\lambda_i b} - e^{\lambda_i a}\right)} dw = \frac{\left(e^{\lambda_i b} - e^{w_1 \lambda_i}\right)}{\left(e^{\lambda_i b} - e^{a \lambda_i}\right)}$. Large values of $\mathbf{P}[S = v_1 | W = w_1]$ increase the likelihood that the result will be published.

The value λ_i is included to scale this probability to be either conservative (i.e., requiring larger test statistics for the investigator to report the result for the exploratory variable ν_i) or liberal (investigators are likely to report exploratory results even if they are only slightly above the lower bound of the critical region.

We may proceed with the application of Bayes Theorem.

$$f_W(w_i | S = \nu_i) = \frac{f_{\upsilon_i}(\nu_i | w_i) f_W(w_i)}{\int\limits_{\Omega_W} f_{\upsilon_i}(\nu_i | w_i) f_W(w_i) dw_i} = \frac{\frac{\lambda_i e^{\lambda_i w_i}}{\left(e^{\lambda b} - e^{\lambda a}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{w_i^{\,2}}{2}} 1_{a \leq w_i \leq b}}{\int\limits_a^b \frac{\lambda_i e^{\lambda_i w_i}}{\left(e^{\lambda_i b} - e^{\lambda_i a}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{w_i^2}{2}} dw_i}$$

$$= \frac{e^{\lambda w} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} 1_{a \leq w \leq b}}{\int\limits_a^b e^{\lambda_i w_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{w_i^2}{2}} dw_i}.$$

By completing the square in the exponent of the integrand, the denominator simplifies to

$$\int\limits_a^b e^{\lambda_i w_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{w_i^2}{2}} dw_i = \frac{1}{\sqrt{2\pi}} \int\limits_0^\infty e^{-\frac{1}{2}\left(w_i^2 - 2\lambda_i w_i + \lambda_i^2 - \lambda_i^2\right)} dw_i$$

$$= e^{\frac{\lambda_i^2}{2}} \int\limits_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(w_i - \lambda_i)^2} dw_i$$

which may be written as

$$e^{\frac{\lambda_i^2}{2}} \mathbf{P}[a \leq N(\lambda_i, 1) \leq b] = e^{\frac{\lambda_i^2}{2}} [\Phi_Z(b - \lambda_i) - \Phi_Z(a - \lambda_i)].$$

Where $\Phi_Z(z)$ is the cumulative distribution function of the standard normal distribution. Thus

$$f_w(w_i | S = v_i) = \frac{e^{\lambda_i w_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{w_i^2}{2}} 1_{a \leq w_i \leq b}}{e^{\frac{\lambda_i^2}{2}} [\Phi_Z(b - \lambda_i) - \Phi_Z(a - \lambda_i)]}.$$

Since

$$p_e(z,b) = \int_{CR_W} f_w(w_i|S=v_i)dw_i = \frac{\int_z^b e^{\lambda_i w_i}\frac{1}{\sqrt{2\pi}}e^{-\frac{w_i^2}{2}}1_{a\le w_i\le b}dw_i}{e^{\frac{\lambda_i^2}{2}}[\Phi_Z(b-\lambda_i)-\Phi_Z(a-\lambda_i)]}$$

we have $p_e(z,b) = \frac{[\Phi_Z(b-\lambda_i)-\Phi_Z(z-\lambda_i)]}{[\Phi_Z(b-\lambda_i)-\Phi_Z(a-\lambda_i)]}$. Recognizing that Φ_Z is a continuous function we can write

$$p_e(z) = \lim_{b\to\infty} p_e(z,b) = \lim_{b\to\infty}\frac{[\Phi_Z(b-\lambda_i)-\Phi_Z(z-\lambda_i)]}{[\Phi_Z(b-\lambda_i)-\Phi_Z(a-\lambda_i)]} = \frac{1-\Phi_Z(z-\lambda_i)}{1-\Phi_Z(a-\lambda_i)}.$$

Confidence intervals.

For a two-sided confidence interval for the exploratory estimate θ estimating a parameter θ with a standard error τ_θ we can compute the value z^*, the lower bound of the test statistic that produced p_e, as $p_e(z^*) = \frac{1-\Phi_Z(z^*-\lambda_i)}{1-\Phi_Z(a-\lambda_i)}$. We find that $z^* = \Phi_z^{-1}(1-\frac{p_e}{2}[1-\Phi_Z(Z_{1-\alpha/2}-\lambda)]) + \lambda$., and we can compute the $1-\frac{p_e}{2}$ size confidence interval for θ, as $(\theta-z^*\tau_\theta,\ \ \theta+z^*\tau_\theta)$.

3. Results

The preceding short derivation provides a straightforward procedure to convert the standard *p*-value produced from exploratory research to a *p*-value that is more reflective of what would be obtained from a confirmatory evaluation. For the following computations, we assume the lower bound of the critical region is $a = 1.96$, and assume $\lambda_i = 1$ for all λ. We then compute the exploratory *p*-value p_e using

$$p_e \quad = \quad \frac{1-\Phi_Z(z-1)}{1-\Phi_Z(a-1)}$$

where *z* is the value of the test statistic from the exploratory analysis, and plot it against the *z*-statistic. The standard *p*-value is also plotted for a one tailed test (Fig. 2).

Fig. 2 reveals that for all test statistics in the critical region, the exploratory *p*-value is substantially larger than the standard *p*-value. In fact, at the lower boundary of the critical region for z = 1.96, while the standard *p*-value is 0.025, $p_e = 1$. As the test statistic *z* increases, the test statistic becomes more persuasive, and the p_e-value falls. In this circumstance, it does not reach 0.05 until approximately $z = 3.38$, demonstrating the substantial signal strength that must be generated by exploratory evaluations for the exploratory finding to be comparable to standard statistical criteria. Fig. 3 demonstrates this relationship as a function of λ, the parameter from the probability density function of the exploratory value given the test statistic. Here, as in Fig. 2, the exploratory test statistic p_e remains high. However, we also see that this relationship can be modified by the choice of λ. For large values of λ, the inflection changes, with even larger values of the exploratory test statistic required before p_e falls below falls into the traditional critical region.

We can also use the methods to compute a confidence interval. Consider a clinical trial which has the exploratory analysis producing an effect size of 9.4 with a standard error of 4. The confidence interval locations and widths for the standard computation and the exploratory assessments for increasing values of λ, are available (Table 1).

In this table, the standard confidence interval is produced simply based on Z = 1.96. Subsequent rows provide the lower and upper bounds of the confidence interval for different values of λ. Note that already for $\lambda = 0$, there is a dramatic increase in the confidence interval width that is produced under the exploratory paradigm. The confidence interval based on the exploratory analysis produces ever larger confidence intervals with larger values of λ.

4. Discussion

The dim view taken by clinical trialists of exploratory analysis is understood and well justified. Beginning at a point in the early history of clinical trials when these types of analyses were considered of the highest value simply because they were generated from a clinical trial, their status plummeted from the 1980's forward as the inability to reproduce their findings eroded faith in their generalizability. It was understandably recognized that exploratory analyses could raise questions but could not be used to answer them.

However, the reason to exclude them was empirical, not theoretical. Since it had been shown repeatedly that exploratory analyses were not generalizable they were deemphasized. No *a priori* theory demonstrated that they were inferior — pragmatism did.

Here we begin a discussion of the specific weaknesses of exploratory analyses from a theoretical basis inaugurating conversations of how their theoretical weaknesses can be repaired. To that end, two

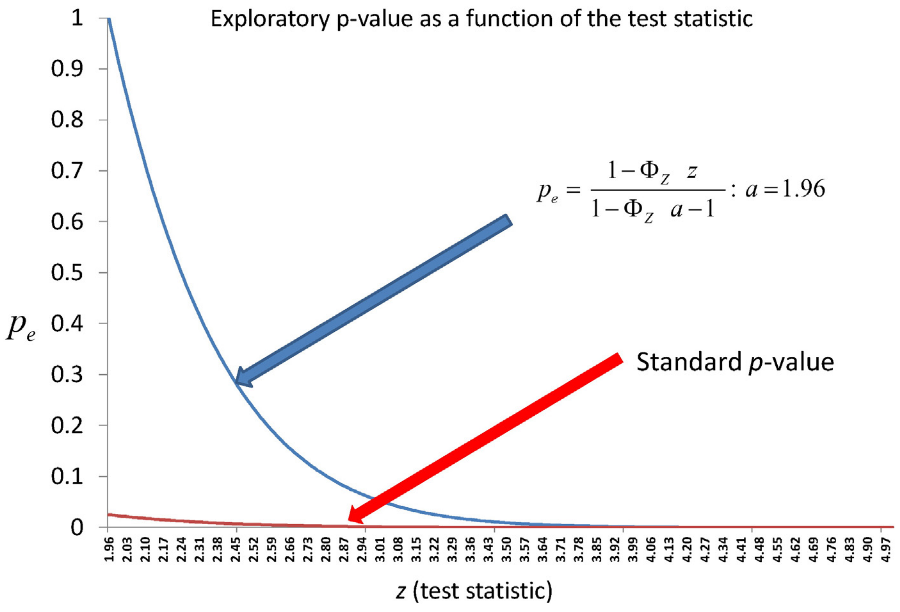

Fig. 2. Exploratory *p*-value as a function of the test statistic. The exploratory *p*-value is much larger than the *p*-value then under the traditional computation.

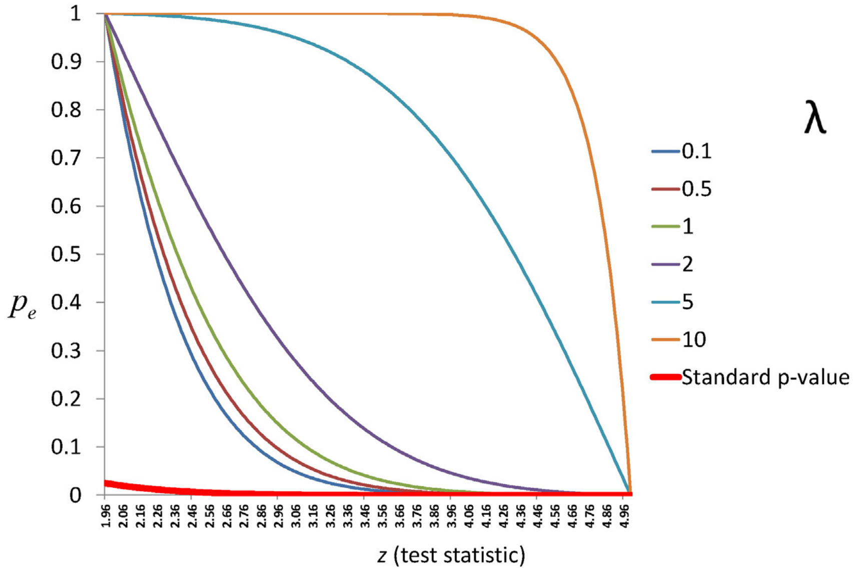

Fig. 3. Exploratory *p*-value as a function of λ. The exploratory *p*-value is much larger than the *p*-value then under the traditional computation.

weaknesses of these analyses have been identified, one practical, and the other — the random selection of the analysis variable — theoretical.

The author could find little prior work on the effect of the random selection of the dependent variable. Random effects models are well developed and widely used [50]. However, what is random in the random effects model is not the variable of analysis, but the levels of explainer variables. Meta analyses use random effects models depicting the random selection of the studies to be incorporated in the evaluation; however, the endpoint variable that has been selected for summarization was not randomly selected. In the standard general linear model, there has been expansion away from the assumption that the explainer variables are fixed [51] but the dependent variable is never selected to be random. Stepwise regression selects explainer variables randomly, but the dependent variable remains fixed, and while authoritative concerns for the stepwise approach have appeared [52] there has been little to no discussion about random selections of the dependent variable. Bayes procedures, while they accept variability in the parameter(s) of the prior distribution, assumes the variable to be analyzed has already been selected [53].

The perspective taken here was to view the theoretical weakness of exploratory analyses through the prism of probability, suggesting that the *p*-value used in exploratory analyses is different from the classic *p*-value used in confirmatory analyses. Both are conditional probabilities — however, the exploratory analysis while it "reverses the given", is interpreted as though the condition was not reversed. The thesis here is that if this conditional probability can be transformed into the conditional probability utilized in confirmatory analyses, the interpretation of exploratory analyses can more closely align with that of confirmatory analyses. The Results section demonstrates that the mathematics required for this transformation is uncomplicated, and the conclusion is that the post-transformation exploratory *p*-value for the same magnitude of the test statistic is substantially larger than the traditional *p*-value. This observation justifies the intuition that findings from exploratory analyses must be overwhelming to be worthy of our attention and it demonstrates how awe-producing such exploratory results must be. In fact one can compute the specific lower bound for the critical region p_e. Confidence intervals are substantially wider reinforcing the notion that the non-prospective nature of exploratory analyses leads to less precise estimators. For many, confidence interval reporting in the hypothesis generating paradigm may be sufficient to convey the results.

Table 1
Effect of λ on 95% confidence interval width effect size of 9.4 with a standard error of 4.

	Standard conf. interval	14.2	20.3
λ	0.1	−3.6	22.4
	0.5	−4.2	23.0
	1	−5.1	23.9
	1.5	−6.2	25.0
	2	−7.5	26.3
	2.5	−9.0	27.8
	3	−10.7	29.5
	3.5	−12.5	31.3
	4	−14.5	33.3
	4.5	−16.4	35.2
	5	−18.4	37.2

The effect of the value of λ on the value of the lower bound of this test statistic is quite clear (Fig. 4). Recalling that λ is the parameter of the exponential density relating the likelihood to report an exploratory result to the size of the test statistic. The value is in the hands of the investigators, and should be chosen prospectively. From Fig. 3, we observe that for each value of λ_i, p_e is close to one for findings that are themselves close to the lower boundary of the critical region. In circumstances where the field is new, and the exploratory work must be corroborated in any event, the procedures developed are unlikely to be of assistance, and exploratory finding will need to be corroborated by a prospectively declared assessment in a future trial. However, if the investigators wish the exploratory analyses to have elevated standing, then they might follow the development outlined here and choose a value of λ_i that is small, on the order of 0.1. However, in a field that is well researched, and the exploratory analyses have been conducted in various formats by other researchers, the recommendation is to choose a larger value of λ_i such as one. For investigators who are likely to report exploratory values even though they are only slightly significant in the standard analysis, λ should perhaps be larger, generating a greater lower bound of the critical region.

As with all theoretical undertakings, the assumptions of the arguments made in this manuscript must be carefully considered. It is assumed that one can place a probability density function on the likelihood that a test statistic will be reported by the investigator given the value of that exploratory analysis' test statistic. This assumption is tenable given the well-recognized tendency of many investigators to report all findings that are significant. However, the choice of the density function itself is an open question. The only way to really be sure of the choice of the density function is to carry out an examination of all of

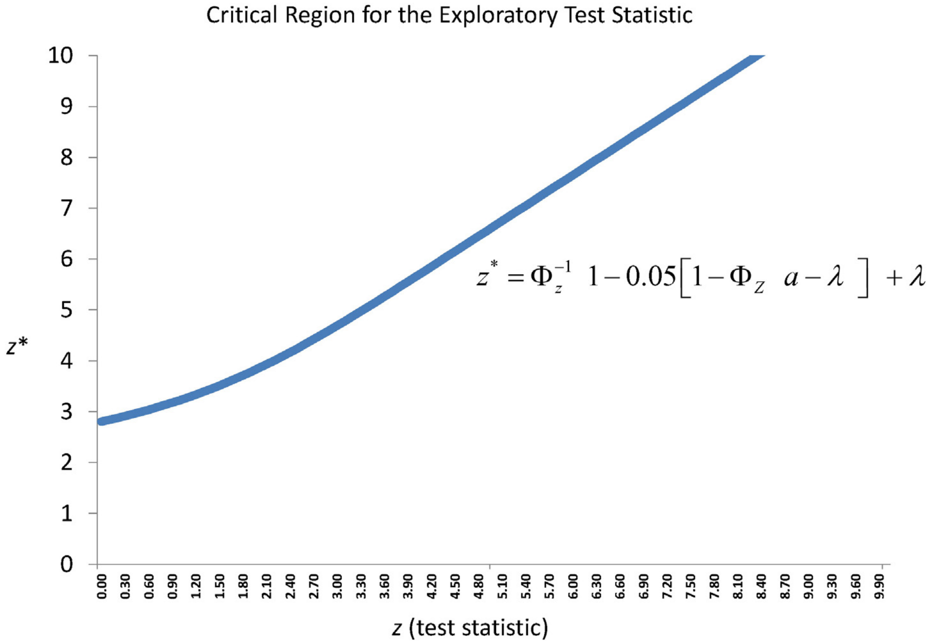

Fig. 4. Lower bound of the five percent critical region as a function of λ. As the propensity to report marginal values of the exploratory finding increases, so does the lower bound of its critical region.

the studies that report exploratory analyses, and then examine the probability distribution of the studies whose exploratory analysis is positive. However, to begin, an educated guess would inform us that a probability density function that increases the likelihood of publishing a result with increasing value of the test statistic such as the one chosen in this manuscript is a reasonable starting point.

However, alternative parameterizations of this density function are certainly admissible and must be examined. For example, one could choose a probability density function related to the logistic function, e.g.,

$$g_{v_i}(v_i|w_i) = \frac{\lambda_i}{(\ln(1+e^{\lambda_i b}) - \ln(1+e^{\lambda_i a}))}\frac{1}{1+e^{-\lambda_i x}}1_{a \le w_i \le b}.$$

The schema used in this manuscript could be applied to $g_{\nu_i}(\nu_i|w_i)$ although the solution would be somewhat more complicated. However, both f (as chosen in the results section) and g have the property of increasing measure assigned to larger values of the test statistic w_i which is consistent with the practice of reporting exploratory findings that are generated by the investigators. Thus, as is the case of the density developed in the body of this manuscript, incorporating a density function that monotonically increases with the test statistic will produce a p_e value that is also large.

Another alternative would be $f_{v_i}(\nu_i|w_i) = \frac{1}{b-a}1_{a \le w_i \le b}$ which assumes that the exploratory analysis will be reported for any value of the test statistic w greater than the lower bound of the critical region. Following the computations elaborated in the results section, we find $p_e = \frac{1-\Phi_Z(z)}{1-\Phi_Z(a-1)}$. The figure relating the relationship of this p-value with the value of the test statistic in the critical region for $a = 1.96$, equivalent to $\lambda = 0$. (Fig. 3), If $a = 1.96$, then $p_e = 40[1 - \Phi_Z(z)]$, or forty times the standard p-value assuming the standard p-value is from a test statistic that is at least 1.96. We can also compute the lower bound of the 0.05 critical region as $z = 3.024$. This would be a useful starting point for working in a new field in which very little work has been conducted. The selections of other density functions would produce different results.

In addition, the ubiquity of the normal distribution as the distribution of the test statistic justifies the assumption here, however other distributions are worthy of consideration (such as binomial, Poisson and chi-square distributions).

One other difficulty offered by exploratory analyses is the multiple testing issue. Left unchecked, the large number of hypothesis tests conducted under the standard exploratory paradigm inflates type I error to unacceptable levels. Alternatively, the simple application of the Bonferroni procedure would exclude many exploratory analyses that might otherwise be worthy of consideration.

An alternative would be to prospectively set aside type I error for exploratory analyses. For example, consider the case of a single clinical trial where there is a simple prospectively declared primary analysis with a familywise alpha level of 0.05, and the expectation that there will be many exploratory evaluations. Permit the investigator to allocate, for example 0.0495 for the primary endpoint that was prospectively declared, leaving 0.005 for the exploratory analyses. The impact on the prospectively declared endpoint analysis would be small, and the effect on the sample size computation would be negligible. An, although the residual 0.005 exploratory alpha is quite low, this in concert with the p_e threshold that the paper develops that already would discard multiple exploratory findings.

The use of Bayes Theorem does not earn the results presented here the appellation "Bayes procedure". There is no prior distribution here, and no loss function is invoked. It is simply the application of conditional probability.

There is likely no quantitative adjustment that will transform or imbue the unplanned, underpowered exploratory analysis with the same features of a well designed and well executed prospective analysis. Just as, in a clinical trial, a small p-value does not adumbrate sloppy execution, here a small critical region does not convert the exploratory analysis into a prospective one. The manuscript proposes that exploratory analyses be identified as early as possible, have adequate resources for data collection, include as much data as possible to increase their power and then apply the p_e value or the exploratory confidence interval. However, for the exploratory analysis which has adequate power and is skillfully executed, the manuscript's methodology can be a useful metric to assess the contribution of the exploratory analysis.

The p-value has been the focus of this discussion, but if history has taught us anything, it is that we cannot rely on them to the exclusion of everything else. Effect sizes, confidence intervals must also be considered.

This realization suggests that an alternative approach to this problem might lie in estimation theory. While the second thought experiment provided earlier seems to suggest that no change in estimation theory is required, alternative views would provide new estimators that function more optimally in the exploratory paradigm, generating new standard errors, and confidence intervals. This is an area worthy of investigation.

Also, this development presumes that we are not required to assess the impact of the random selection of the analysis variable on the estimator's formula. The properties of unbiasedness and minimum variance are essential features of estimators when experiments are repeated. However our familiar estimators lose these features when one is attempting to reproduce the experiment of first randomly choose a variable then randomly choose a sample. If this is determined to be a worthy goal as well, then additional work in estimation theory is required.

Finally, we must return to the first articulated weakness of exploratory analyses — the logistical consideration. It is difficult to imagine that any theoretical elaboration, no matter its elegance will overcome poor data collection procedures. This recognition permits advice to the investigators expecting that exploratory analyses to be conducted at the trial's end.

Investigators must first commit to the collection of precise data in their study. Even if data (e.g., laboratory data, quality of life data, biomarker data) is not prospectively declared, the investigators can and should commit themselves to high precision in data collection. Since data is collected with the presumption it will be analyzed, its future analysis requires present and explicit quality control. The ethics of imprecise measures of human data in clinical trials requires this.

Much additional work is still required to attain a solid solution to the exploratory analyses dilemma. Once the precision of the data is ensured, and λ is chosen prospectively, the simple analysis tool for *p*-value interpretation generated here can be applied. The implication is that many exploratory evaluations that reach nominal significance will not fall into the critical regions provided here. However, the few that do perhaps should have the constraining moniker "exploratory" if not removed, then relaxed.

Funding

Partial funding for this investigator provided by the National Heart, Lung, and Blood Institute under cooperative agreement UM1 HL087318.

References

[1] E.J. Emanuel, The future of biomedical research, JAMA 309 (2013) 1589–1590.
[2] M.S. Lauer, Personal reflections on big science, small science, or the right mix, Circ. Res. 114 (2014) 1080–1082.
[3] M. Rosbash, A threat to medical innovation, Science 333 (2011) 136.
[4] M. Hanna, Matching taxpayer funding to population health needs, Circ. Res. 116 (2015) 1296–1300.
[5] M.S. Lauer, D. Gordon, M. Olive, Matching taxpayer funding to population health needs: not so simple, Circ. Res. 116 (2015) 1301–1303.
[6] R.T. O'Neill, Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance, Control. Clin. Trials 18 (1997) (550,6; discussion 561-7).
[7] S.J. Pocock, N.L. Geller, A.A. Tsiatis, The analysis of multiple endpoints in clinical trials, Biometrics 43 (1987) 487–498.
[8] Anonymous, Food and Drug Administration Amendments Act of 2007, 2007 110–185.
[9] E.A. Gehan, N.A. Lemak, Statistics in Medical Research: Developments in Clinical Trials, Plenum Medical Book Company, New York, 1994.
[10] A.B. Hill, Observation and experiment, N. Engl. J. Med. 248 (1953) 995–1001.
[11] P. Armitage, Bradford Hill and the randomized controlled trial, Pharmaceutical 6 (1992) 23–37.
[12] R.A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, Edinburg, 1925.
[13] R.A. Fisher, The arrangement of field experiments, J. Minist. Agric. (1926) 503–513 (September).
[14] J. Berkson, Tests of significance considered as evidence, J. Am. Stat. Assoc. 37 (1942) 335–345.
[15] R.A. Fisher, Response to Berkson, J. Am. Stat. Assoc. 37 (1942) 103–104.
[16] J. Berkson, Experiences with tests of significance. A reply to R.A. Fisher, J. Am. Stat. Assoc. 37 (1942) 242–246.
[17] S.N. Goodman, Toward evidence-based medical statistics. 1: the P value fallacy, Ann. Intern. Med. 130 (1999) 995–1004.
[18] A.M. Walker, Significance tests represent consensus and standard practice, Am. J. Public Health 76 (1986) 1033–1034.
[19] J.L. Fleiss, Significance tests have a role in epidemiologic research: reactions to A. M. Walker, Am. J. Public Health 76 (1986) 559–560.
[20] J.L. Fleiss, Confidence intervals vs significance tests: quantitative interpretation, Am. J. Public Health 76 (1986) 587–588.
[21] J.L. Fleiss, Dr. Fleiss response, Am. J. Public Health 76 (1986) 1033–1034.
[22] A.M. Walker, Reporting the results of epidemiologic studies, Am. J. Public Health 76 (1986) 556–558.
[23] R. Peto, M.C. Pike, P. Armitage, N.E. Breslow, D.R. Cox, S.V. Howard, et al., Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design, Br. J. Cancer 34 (1976) 585–612.
[24] R. Peto, M.C. Pike, P. Armitage, N.E. Breslow, D.R. Cox, S.V. Howard, et al., Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples, Br. J. Cancer 35 (1977) 1–39.
[25] Anonymous, Multiple risk factor intervention trial risk factor changes and mortality results. Multiple Risk Factor Intervention Trial Research Group, JAMA 248 (1982) 1465–1477.
[26] L.H. Kuller, S.B. Hulley, J.D. Cohen, J. Neaton, Unexpected effects of treating hypertension in men with electrocardiographic abnormalities: a critical analysis, Circulation 73 (1986) 114–123.
[27] Anonymous, The 1984 report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure, Arch. Intern. Med. 144 (1984) 1045–1057.
[28] A.M. Feldman, M.R. Bristow, W.W. Parmley, P.E. Carson, C.J. Pepine, E.M. Gilbert, et al., Effects of vesnarinone on morbidity and mortality in patients with heart failure. Vesnarinone Study Group, N. Engl. J. Med. 329 (1993) 149–155.
[29] J.N. Cohn, S.O. Goldstein, B.H. Greenberg, B.H. Lorell, R.C. Bourge, B.E. Jaski, et al., A dose-dependent increase in mortality with vesnarinone among patients with severe heart failure. Vesnarinone Trial Investigators, N. Engl. J. Med. 339 (1998) 1810–1816.
[30] B. Pitt, R. Segal, F.A. Martinez, G. Meurers, A.J. Cowley, I. Thomas, et al., Randomised trial of losartan versus captopril in patients over 65 with heart failure (Evaluation of Losartan in the Elderly Study, ELITE), Lancet 349 (1997) 747–752.
[31] B. Pitt, P.A. Poole-Wilson, R. Segal, F.A. Martinez, K. Dickstein, A.J. Camm, et al., Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: randomised trial—the Losartan Heart Failure Survival Study ELITE II, Lancet 355 (2000) 1582–1587.
[32] M. Packer, C.M. O'Connor, J.K. Ghali, M.L. Pressler, P.E. Carson, R.N. Belkin, et al., Effect of amlodipine on morbidity and mortality in severe chronic heart failure. Prospective Randomized Amlodipine Survival Evaluation Study Group, N. Engl. J. Med. 335 (1996) 1107–1114.
[33] M. Packer, Presentation of the Results of the Prospective Randomized Amlodipine Survival Evaluation-2 Trial (PRAISE-2), 2000.
[34] M.A. Pfeffer, The Second Prospective Randomized Amlopdipine Survival Evaluation (PRAISE-2), 2000.
[35] M.A. Pfeffer, H. Skali, PRAISE (prospective randomized amlodipine survival evaluation) and criticism, JACC Heart Fail. 1 (2013) 315–317.
[36] M. Packer, M.R. Bristow, J.N. Cohn, W.S. Colucci, M.B. Fowler, E.M. Gilbert, et al., The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. U.S. Carvedilol Heart Failure Study Group, N. Engl. J. Med. 334 (1996) 1349–1355.
[37] L.A. Moyé, D. Abernethy, K. von Olshausen, T. Pop, J. Berger, M. Packer, et al., Carvedilol in patients with chronic heart failure, N. Engl. J. Med. 335 (1996) 1318–1320.
[38] M. Packer, J.N. Cohn, W.S. Colucci, Response to Moyé and Abernethy, N. Engl. J. Med. 335 (1996) 1318–1319.
[39] L.D. Fisher, Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing, Control. Clin. Trials 20 (1999) 16–39.
[40] L.D. Fisher, L.A. Moye, Carvedilol and the Food and Drug Administration approval process: an introduction, Control. Clin. Trials 20 (1999) 1–15.
[41] L.A. Moye, End-point interpretation in clinical trials: the case for discipline, Control. Clin. Trials 20 (1999) (40,9; discussion 50-1).
[42] L.D. Fisher, Carvedilol and the Food and Drug Administration—approval process: a brief response to Professor Moye's article, Control. Clin. Trials 20 (1999) 50–51.
[43] K.J. Rothman, S. Greenland, T.L. Lash, Modern Epidemiology, 3rd ed. Lippincott, Williams, and Wilkins, Philadelphia, 2012.
[44] D. LeRoith, S.I. Taylor, J.M. Olefsky, Chapter 6 — sulfonylurea receptors, ATP-sensitive potassium channels, and insulin secretion, Diabetes Mellitus: A Fundamental and Clinical Text, Lipopincott, Williams, and Wilkins, Philadelphia, 2000.
[45] U.S. Anonymous, Department of Health and Human Services. FDA. CDER, CBER, Guidance for industry E9 statistical principles for Clin. Trials (1998).
[46] N. Freemantle, Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? BMJ 322 (2001) 989–991.
[47] L.A. Moye, Statistical Monitoring of Clinical Research: Fundamentals for Investigators, Springer, New York, 2005.
[48] L.A. Moye, Statistical Reasoning in Medicine - The P value Primer, 2nd ed. Springer, New York, 2006.
[49] K.K. Lan, J. Wittes, The B-value: a tool for monitoring data, Biometrics 44 (1988) 579–585.
[50] R. Christensen, Plane Answers to Complex Questions: The Theory of LInear Models, Third ed. Springer, New York, 2002.
[51] M. Kutner, C. Nachtsheim, J. Neter, Applied Linear Statistical Models, 5th ed. McGraw-Hill Irwin, New York, 2004.
[52] F.E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer-Verlag, New York, 2001.
[53] J. Berger, Statistical Decision Theory and Bayesian Analyses, 2nd ed. Springer-Verlag, New York, 1985.