Bias and Conditioning in Sequential Medical Trials

Author(s): Cecilia Nardini and Jan Sprenger

Source: *Philosophy of Science*, Vol. 80, No. 5 (December 2013), pp. 1053-1064

Published by: The University of Chicago Press on behalf of the Philosophy of Science Association

Stable URL: https://www.jstor.org/stable/10.1086/673732

Bias and Conditioning in Sequential Medical Trials

Cecilia Nardini and Jan Sprenger*†

Randomized controlled trials are currently the gold standard within evidence-based medicine. Usually they are monitored for early signs of effectiveness or harm. However, evidence from trials stopped early is often charged with bias toward implausibly large effects. To our mind, this skeptical attitude is unfounded and caused by the failure to perform appropriate conditioning in the statistical analysis of the evidence. We contend that conditional hypothesis tests give a superior appreciation of the obtained evidence and significantly improve the practice of sequential medical trials, while staying firmly rooted in frequentist methodology.

1. Introduction. Randomized controlled trials (RCTs)—trials where patients are randomly assigned to a treatment and a control group, while controlling for possible confounders—are currently the gold standard within evidence-based medicine (Worrall 2007). Usually they are conducted as sequential trials allowing for monitoring for early signs of effectiveness or harm.

In sequential trials, data are typically monitored as they accumulate. That is, we have interim looks at the data, and we may decide to stop the trial before the planned sample size is reached. By terminating a trial when overwhelming evidence for the effectiveness or harmfulness of a new drug is

*To contact the authors, please write to: Cecilia Nardini, University of Milan and European Institute of Oncology (IEO), Campus IFOM-IEO, Via Adamello, 16, 20139 Milan, Italy; e-mail: nardini.folsatec@gmail.com. Jan Sprenger, Tilburg Center for Logic and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: j.sprenger@uvt.nl.

†The authors would like to thank the senior and junior members of the FOLSATEC PhD program, as well as David Teira and the PSA audience. Jan Sprenger would also like to thank the Netherlands Organization for Scientific Research (NWO) for supporting this research through Veni grant no. 016.104.079.

Philosophy of Science, 80 (December 2013) pp. 1053–1064. 0031-8248/2013/8005-0036$10.00

available, we can bound the prohibitive costs of a medical trial and protect in-trial patients against receiving inferior treatments. Thus, monitoring contributes to meeting ethical and epistemic requirements that clinical investigators are confronted with.

However, the early termination of sequential trials raises an important ethical concern: Is it mandatory to stop a trial as soon as the new treatment shows convincing signs of superiority? Or should the trial be continued in order to achieve a result that would convince the wide medical community of the superiority of the new treatment? On the one hand, the health of actual patients must not be jeopardized by administering an inferior treatment; on the other hand, establishing sound and univocal scientific conclusions will facilitate an effective cure of future patients.

The issue is complicated by the fact that evidence from trials stopped early is often met with skepticism in the medical literature: "RCTs stopped early for benefit . . . show implausibly large treatment effects. . . . Clinicians should view the results of such trials with skepticism" (Montori et al. 2005, 2203). This standpoint is affirmed by the recent STOPIT-2 metastudy where Bassler et al. (2010, 1187) blame truncated RCTs for "appreciable overestimates of effect."

While we cannot adjudicate the far-reaching question about the ethical legitimacy of monitoring, we side with Worrall (2008, 418) that "no informed view of the ethical issues . . . can be adopted without first taking an informed view of the evidential-epistemological ones." In particular, we think that the skeptical attitude about trials stopped early for benefit stems from a fallacious statistical interpretation of such trials. These misinterpretations are, to our mind, mainly caused by a lack of awareness about issues in statistical methodology that also troubles other disciplines, such as economics and psychology. Indeed, the two grand schools of statistical inference—Bayesian and frequentist inference—are in outright conflict about how to plan and to evaluate a sequential trial.

Our essay takes the following route. First, we expose the arguments for and against the presence of bias in early stopped trials and explain why this problem is related to principled questions in statistical methodology (sec. 2). Subsequently, we argue that the real problem is the use of unconditional error assessments in sequential trials, rather than the often-invoked divide between Bayesians and frequentists (sec. 3). Then we show that conditional frequentist tests reconcile the need for valid postexperimental appraisal of the evidence with the realities of the current regulatory framework in medicine and, in particular, with the implied preference for frequentist analysis (sec. 4). Finally, we wrap up our results and sketch how a superior methodological framework can improve the design and practice of sequential trials and eventually lead to better decisions (sec. 5).

2. The Assessment of Truncated Trials. The practice of stopping RCTs early for benefit has been subject to severe epistemological criticism. Skepticism surrounds the results of these trials, due to the fact that they show implausibly large treatment effects, relative to what the medical community would be inclined to expect. In a review of 134 trials stopped early for benefit, Montori et al. (2005) point to an inverse correlation between sample size and treatment effect: the smaller the sample size achieved by the trial at the moment of stopping, the larger the estimate it provided for the effect. These findings are supported by a more recent study by Bassler et al. (2010), where truncated trials report significantly higher effects than trials that were not stopped early.

Some prominent cases seem to corroborate this skepticism. Mueller et al. (2007) report a case of two leukemia treatments where interim analyses suggested a high relative risk reduction (53% and 45%) in a particular chemotherapy regimen. However, that assessment had to be reversed after completion of the trial. In the medical community, such cases fuel mistrust toward anticipated claims of benefit and nourish the fear of promoting a treatment that is actually less efficacious. Therefore, stopping a trial early might lead to a result that the medical community does not trust, canceling the epistemic and ethical benefits that monitoring possesses in the long run.

However, not all methodologists share this pessimistic view on trials stopped early. Goodman, Berry, and Wittes (2010) observe that pronounced effect size differences between truncated and completed trials are actually predictable: highly efficacious treatments will naturally be more prone to early termination for benefit. Hence, the observed difference in estimated effect size is precisely what we should expect. Comparing truncated to completed trials amounts, as highlighted by Berry, Carlin, and Connor (2010), to selecting the trials to be compared on the basis of their outcome.

In this context, prior knowledge or empirically based prior expectations are highly relevant for sound decision making. Unfortunately, at present they enter the final decisions only in a methodologically unsatisfactory ad hoc way. This observation suggests that systematic use of Bayesian inference may address the problem. A Bayesian represents subjective uncertainty by means of a prior probability distribution over the values of the quantity of interest (e.g., relative risk reduction). By means of Bayes's theorem, this distribution is updated to a posterior probability distribution that synthesizes the observed evidence with the background knowledge.

In the Bayesian framework, implausibly large observed effects can be balanced by prior expectations and lead to a more conservative conclusion than in standard frequentist methodology. In particular, it can be explained that truncated trials provide, ceteris paribus, less confidence than trials with

a comparable effect size that were completed (Goodman 2007). The smaller the actual sample, the more will the posterior distribution resemble the prior distribution, for a given effect size. So it appears that the worries of Montori et al. (2005) and Bassler et al. (2010)—overestimation of treatment effect in truncated RCTs—could be alleviated by switching the statistical framework.

Despite the advantages just outlined, there are some serious counterarguments to the viability of Bayesianism in clinical trials. First of all, the specification of a prior probability function (and a decision model) is problematic in a number of ways (see Moyé 2008). Second, in Bayesian statistics, experimental design is apparently irrelevant for the postexperimental conclusions. This is unacceptable to regulatory bodies that are keen to promote proper design of medical trials as a means to ensure the validity of trial results (e.g., Food and Drug Administration 2010).

Even though some of these worries are regulatory rather than epistemological, they are certainly legitimate. Indeed, we believe that solving the interpretational problems with truncated trials does not require one to pass from the frequentist to the Bayesian paradigm. As we will argue in the upcoming sections, it is more fruitful to turn to a different distinction: namely, to replace unconditional by conditional procedures.

3. Problems with Unconditional Inference in Sequential Medical Trials. Sequential medical trials usually control the reliability of a testing procedure from a preexperimental point of view, by means of type I and type II error rates. These error probabilities are extremely important for proper experimental design, and they get a lot of attention from a regulatory point of view. Moreover, Mayo and Kruse (2001) have argued, among others, that if the sampling plan is violated, error probabilities cannot be properly controlled and are actually inflated far beyond acceptable.

However, adherence to a proper sequential sampling plan is not sufficient to secure a reliable result. Arguably, what is most disturbing to the medical community is the fact that, according to current procedures, a truncated trial has prima facie the same reliability as a trial carried to the planned end. This is because Neyman and Pearson's type I and II error rates are unconditional quantities; that is, they are insensitive to whether the data are just at the significance boundary or far beyond it.

In line with this observation, we contend that the unconditional nature of Neyman-Pearson hypothesis tests is the culprit for their epistemological shortcomings. To motivate this claim, we walk the reader through an example by Cox (1958) and Royall (1997, 74–75). Suppose that we test $H_0: \mathcal{N}(0, \sigma^2)$ against $H_1: \mathcal{N}(1, \sigma^2)$ with known σ^2 and that the toss of a fair coin decides whether we draw $N = 1$ or $N = 100$ independently and identically distributed observations. It seems natural to apply the most powerful test at

the 5% level in either case. However, the probabilistic mixture of the two most powerful tests at the 5% level is not the most powerful test in the overall experiment. We can do better if we reject H_0 for $x_1 > 1.282$ in the case of $N = 1$, while rejecting H_0 if $\bar{x} > 0.508$ in the case of $N = 100$. Both procedures are tests at the 5% level, but the second, "gerrymandered" test has a greater power (69%) than the mixture of unconditional tests (63%).

One may be inclined to dismiss the second test because not all of its components are tests at the 5% level. In the $N = 1$ case, the nominal significance level of the test is 10%. However, from an unconditionalist (preexperimental) viewpoint, only the overall error rates should count. Here, the superior power features speak for the second, gerrymandered test. This feature of the prevalent Neyman-Pearson methodology reveals the tension between the preexperimental design of unconditional procedures and the need to efficiently learn from the actual data. Unconditional error rates and confidence intervals do not address that second goal: "Now if the object of the analysis is to make statements by a rule with certain specified long-run properties, the unconditional test . . . is in order. . . . If, however, our objective is to say what we can learn from the data we have, the unconditional test is surely no good" (Cox 1958, 360). The example can, of course, be easily generalized. It undermines the view that unconditional, preexperimental error probabilities can qualify the goodness of an inference.

Therefore, practitioners who rely on unconditional procedures have to find informative and reliable postdata assessments of the evidence. Often, they report the observed p-value to quantify the conclusiveness of the rejection of the null. However, p-values really combine the worst of all worlds. Since comprehensive and devastating criticisms of using p-values in scientific experiments have been delivered elsewhere (Royall 1997; Goodman 1999); we only mention their most fundamental failures: they neither possess a valid frequency interpretation nor do they provide a useful measure of confidence in the null hypothesis.

Moving to confidence intervals is often suggested as a way of circumventing the p-value problem (e.g., Cumming and Finch 2005). However, a 95% confidence interval merely specifies the set of parameter values that are consistent with the observation at the 95% level. This does not mean that we should have 95% confidence that the confidence interval includes the parameter value. In fact, the degree of confidence is just an average coverage rate over intervals from repeated random samples; it is not the coverage probability of the one particular interval that the investigator happens to get. Therefore, some confidence intervals may include the entire sample space (see Seidenfeld 1981), raising the question of whether the entire notion is a misnomer.

These problems of unconditional inference can be overcome by conditioning on the relevant chunks of information. In the next section, we will

see how conditional inference may resolve the methodological confusion about interpreting truncated RCTs without abandoning the framework of frequentist statistics.

4. Conditional Frequentist Inference. Conditional inference tries to improve upon unconditional procedures by quantifying the degree of confidence that we can have in our conclusions as a function of the observed evidence. More precisely, conditional inference builds on the strength of the observed evidence. As we will show in this section, it can be justified from both the Bayesian and the frequentist perspective. The idea comes up for the first time in Cox's (1958) seminal paper and has been developed later by Kiefer (1977) and Berger (2003), together with various coauthors.

The main idea can be motivated by a very simple example (Kiefer 1977; Berger 2003). Two observations X_1 and X_2 are taken with probability law

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2. \end{cases}$$

If we now construct a confidence interval for θ, then the interval $C_\theta(\cdot,\cdot)$ defined by

$$C_\theta(X_1, X_2) := \begin{cases} X_1 + 1 & \text{if } X_1 = X_2 \\ (X_1 + X_2)/2 & \text{if } X_1 \neq X_2 \end{cases}$$

has an unconditional coverage of 75%. Yet this does not seem to be a sensible conclusion regarding the confidence that the data warrant with respect to the true value of θ. Dependent on whether we observe $|X_1 - X_2| = 0$ or $|X_1 - X_2| = 2$, we are entitled to a statement with (a posteriori) confidence 50% and 100%, respectively. The unconditional coverage of 75% neglects that, after learning the strength of the evidence (i.e., the value of $|X_1 - X_2|$), we are in a much better position to assess the confidence which the data grant about our inference. Thus, conditioning on the value of $|X_1 - X_2|$ improves the accuracy of our conclusions (see Cox 1958, 361–63).

It is also noteworthy that the probability distribution of $|X_1 - X_2|$ does not depend on the value of θ. That is, $|X_1 - X_2|$ is an ancillary statistic with regard to θ. In particular, conditioning on the value of $|X_1 - X_2|$ is quite different from Bayesian conditionalization: where Bayesians change their subjective probability distributions by conditioning on the entire data, conditioning on the value of $|X_1 - X_2|$ just helps to better appreciate the (frequentist) interpretation of the data.

If this idea is applied to hypothesis testing, which is the major issue in medical trials, unconditional error rates are replaced by conditional error probabilities. In the following, we will outline the basic idea of conditional tests, following Berger, Brown, and Wolpert (1994).

Consider, for the purpose of mathematical convenience, the case of testing a point null hypothesis $H_0: \theta = \theta_0$ against the simple alternative $H_1: \theta = \theta_1$ in some probability model $(\mathcal{X}\, \mathcal{B}(\mathcal{X});\, \theta \in \Theta)$. Define $f_0(x)$ and $f_1(x)$ as the probability densities of data $x \in \mathcal{X}$ under the hypotheses H_0 and H_1, and let the Bayes factor $B(x) := f_0(x)/f_1(x)$ be the ratio of the probability density functions. Now, let F_0 and F_1 be the cumulative distribution functions corresponding to the Bayes Factor.

$$F_0(c) := P_{H_0}(B(X) \leq c) \quad F_1(c) := P_{H_1}(B(X) \leq c).$$

We now divide $\mathcal{X}$ into a partition $(\mathcal{X}_s)_{s \in [0,1]}$ defined by

$$\mathcal{X}_s := \{x \in \mathcal{X} | B(x) = s \vee B(x) = F_0^{-1}(1 - F_1(s))\}. \tag{1}$$

The different $\mathcal{X}_s$ represent, intuitively, different observed strengths of evidence. This can also be made precise mathematically: under the assumption $F_0(1) = 1 - F_1(1)$, which is satisfied for many distributions used in practice, the $\mathcal{X}_s$ have the same probability density under H_0 and H_1, for all values of s. In other words, their distribution is independent of which hypothesis is true (Berger et al. 1994, 1789–90).

This ancillarity property is shared with the statistic $|X_1 - X_2|$ in the above toy example. Therefore, $\mathcal{X}_s$ is excellently suited for the purpose of conditioning: to take the observed strength of the evidence into account without already telling us something about the parameter of interest. Thus, conditioning exploits a crucial strength of Bayesian paradigm—to identify a sensible measure of evidence—without assigning a subjective probability to competing hypotheses.

The conditional error probability can now be calculated by conditioning on the particular set $\mathcal{X}_s$ in which the observed data fall. In particular, we can define a conditional frequentist test by

$$T^*(X) = \begin{cases} \text{reject } H_0 & \text{if } B(X) < 1 \\ \text{accept } H_0 & \text{if } B(X) \geq 1 \end{cases}, \tag{2}$$

and for observed $B(x) = s$, we report conditional error probabilities

$$\alpha(s) = P_{H_0}(\text{reject } H_0 \mid X \in \mathcal{X}_s) = \frac{s}{1+s}, \tag{3}$$

$$\beta(s) = P_{H_1}(\text{accept } H_0 | X \in \mathcal{X}_s) = \frac{1}{1+s}, \tag{4}$$

where the latter equalities have been proven by Berger et al. (1994, theorem 1). Clearly, by using the conditional instead of the unconditional error

probabilities, we gain a much better appreciation of the chance of a wrong decision, given the particular data that we have observed. The higher the Bayes factor, the more confident we can be about an acceptance of the null and vice versa. In particular, the classical, unconditional test just detects whether the data are within or outside the rejection region (and leaves the rest to the notorious *p*-values) whereas the conditional test allows for a fine-grained, properly frequentist discrimination among trials with significant outcomes.

We turn now to briefly discussing a couple of objections that could be made from within the frequentist perspective. First, it could be argued that T^* makes it far too easy to reject the null ($B(X) < 1$) whereas in medicine, evidence has to be really strong before we are convinced of the efficacy of a new treatment and approve of the drug. To this we simply respond that T^* has been selected because of its simplicity, but we can easily change the rejection region according to contextual requirements. To obtain a sensible conditional test, we will often have to use nonancillary conditioning statistics and to include a no-decision region (Berger, Boukai, and Wang 1997, 145–47). However, these features align well with the caution toward premature conclusions that prevails in the medical community and do not pose any problem for the practitioner.

Second, there may be worries about the scope of the above procedure, which we have only explained for the easiest possible case of hypothesis testing. However, Berger et al. (1997) have extended conditional tests to simple versus composite testing problems and, in particular, to the two-sided null hypothesis testing problems that frequently occur in RCTs.

Third, the use of the Bayes factor may indicate that the conditional test is actually a Bayesian test in frequentist cloths. Indeed, for impartial priors $p(H_0) = p(H_1) = 1/2$ the posteriors

$$p(H_0|x) = \left[1 + B(x)^{-1}\right]^{-1} = \frac{B(x)}{1 + B(x)},$$

$$p(H_1|x) = \left[1 + B(x)\right]^{-1} = \frac{1}{1 + B(x)}$$

just correspond to the conditional error probabilities for rejecting and accepting H_0, respectively. However, $B(X)$ possesses a frequentist interpretation, too, since it identifies the most powerful frequentist test in the simple versus simple testing problem.[1]

1. This is the content of the Neyman-Pearson lemma. Furthermore, Berger (2003) introduced a conditional test that relies on the *p*-value as the conditioning statistics and yields the same postdata error probabilities as T^*.

Thus, Bayesians and frequentists can conduct the same (conditional) test and obtain the same numerical conclusions. For the medical practitioner, philosophical questions about the interpretation of probability are clearly secondary as long as there is methodological agreement on procedures and postexperimental data assessment (see Berger 2003). In this sense, conditional inference is a genuine reconciliation of Bayesian and frequentist methodology and a real asset for practitioners.

We would like to conclude this section by means of an application of conditional inference to sequential medical trials. The example involves a trial for adjuvant therapy in resectable hepatocellular carcinoma (Lau et al. 1999). The trial was stopped early based on interim findings, but additional data were available after the decision to stop was made. Pocock and White (1999) describe the situation in detail:

> At the planned interim analysis, the local disease recurrence rates for the active treatment (intra-arterial lipiodol-iodine-131) and control (no adjuvant treatment) groups were three/14 (21%) and 11/16 (69%) respectively ($p = 0.01$). According to the predefined stopping rule, $p < 0.029$ was sufficient for early stopping. . . . Thus, the investigators decided to stop the trial. [However], 13 more patients were randomised before the trial was stopped, and the investigators also decided to postpone analysis while patients already randomised were followed up. Hence, the report (18 months after the trial was stopped) reveals updated recurrence rates of six/21 (29%) and 13/22 (59%), respectively ($p = 0.04$). Thus the absolute difference in recurrence rates shrank from 48% to 30% during the interval between stoppage and publication. (1999, 944)

Such shrinkage of the estimated benefit between the interim and the final analysis is precisely what fuels clinicians' worries about "stopping on a random high" and adds to their skepticism about truncated trials. In this situation conditional error rates can provide real guidance. We set up an alternative hypothesis H_1 according to Lau et al.'s (1999) expectations that "131I-lipiodol would reduce the rate of recurrence [postulated to be 50%] by 50% and double the disease-free survival rate" (1999, 798).

Using this value in calculation of the Bayes factor $B(x) = 0.09$ yields a conditional type I error rate of $\alpha^* = 9\%$ at the interim analysis, instead of the unconditional error rate of $\alpha = 5\%$.[2] Moreover, we can dismiss the apparently strong unconditional p-value of $p = .01\%$, which is just indicative of an unexpectedly high performance. By contrast, the conditional error reflects the greater statistical uncertainty associated with the small

2. Since the trial was stopped following a proper group sequential rule, α remains the same regardless of when the trial is terminated, unlike in Wald's (1947) classical sequential probability ratio test.

sample when the decision to stop the trial was made. At the end of the trial, the conditional test still rejects the null, but the probability of error is now higher: the calculation based on $B(x) = 0.16$ yields a 14% probability of error, which is in line with the reservations of the clinicians involved.

We now briefly wrap up the advantages of conditional over unconditional inference. First, the assessment of the error probability depends on the observed data and is thus way more informative than in the unconditional framework. This alleviates the interpretational problem mentioned in section 2, since conditional error allows medical readers to assess the confidence in the outcome based on the observed data. Clearly, medical investigators should be more concerned with the actual probability of drawing the wrong inference than with the absolute (unconditional) error rate of the testing procedure, also because clinicians have to make ethical decisions for their actual patients (see Nardini 2013).

As a further point, the error probabilities (3) and (4) are independent of the stopping rule, that is the sampling plan determining when the trial is terminated. In an RCT, the stopping rule can never be fully specified, since one cannot cover in advance all eventualities that might happen during a sequential trial. Independence from the stopping rule entails that interpretation of the results and assessment of error are possible even if the stopping rule was misspecified or could not be adhered to due to unforeseen circumstances. This is a substantial practical asset (see Sprenger 2009).

This should not be misunderstood as the claim that predata analysis and experimental design are superfluous. Unfortunately, Berger et al. (1994, 1803) make a claim in that direction, but given the strong emphasis on careful design by methodologists and regulatory bodies (see Moyé 2008; Food and Drug Administration 2010), this is unlikely to increase the acceptance of the conditional approach among medical practitioners. We would like to stress that no such claim is required for making a case for the superiority of the conditional frequentist approach. Moreover, since conditional tests can be conducted from both a Bayesian and a frequentist perspective, practitioners do not have to decide for either camp.

Finally, there are interesting implications for the philosophy of statistics: if the "error statisticians" (e.g., Mayo 1996) are right that learning from error is indeed a cornerstone of inductive inference, then a move to conditional inference may protect their framework against the objections that we have mentioned in section 3. In particular, there is no need to tie an error-statistical methodology to unconditional inference. However, further developing this line of thought goes beyond the scope of this article.

5. Conclusions. In this article, we have analyzed the impact of statistical methodology on a substantive ethical and societal question, namely, data

monitoring in sequential medical trials. In the medical literature, trials stopped early for benefit are often charged with being biased toward implausibly large treatment effects (e.g., Bassler et al. 2010).

We think that this worry is based upon a misinterpretation of sequential trials that is in turn due to shortcomings of standard frequentist procedures. It has been argued (e.g., Goodman 2007) that a Bayesian perspective overcomes this problem: if a trial is stopped early because of an implausibly large effect, blending its result with a (conservative) prior probability distribution naturally mitigates the conclusion. However, as a matter of research tradition and regulatory requirements—in particular, concerns about individual biases in generating prior distributions—the Bayesian framework does not provide an easy way out.

In this essay, we contend that the real issue is not the contrast between Bayesian and frequentist methodology. Rather, we are concerned about the shortcomings of unconditional inference. We have elaborated that while unconditional error probabilities may be helpful in the design of an experiment, they do not tell us what we have actually learned from the data. We have therefore defended proper conditioning—calculating error probabilities conditional on the strength of the observed evidence—as a way of curing the deficits of unconditional frequentist inference. This approach has a natural application to sequential testing and both a valid Bayesian and a valid frequentist interpretation.

As we have demonstrated in a brief example, this approach holds considerable promise for the interpretation of early stopped trials in medicine. The possibility of postdata assessments of the probability of an erroneous conclusion represents an invaluable asset for the practitioner and the decision maker. The results of a medical trial tell much more than the simple acceptance or rejection of a scientific hypothesis: they indicate where evidence is strong and where it is inconclusive, indicating the need for further research. Conditional inference, we believe, can improve the methodology of clinical trials because it allows us to take this additional information into account. In conclusion, a clearer view on issues in statistical methodology can help to better appreciate data from sequential medical trials and lead to more efficient and ethically superior decisions in medical research.

REFERENCES

Bassler, Dirk, et al. 2010. "Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects." *Journal of the American Medical Association* 303:1180–87.

Berger, James O. 2003. "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" *Statistical Science* 18:1–12.

Berger, James O., Ben Boukai, and Yinping Wang. 1997. "Unified Frequentist and Bayesian Testing of a Precise Hypothesis." *Statistical Science* 12:133–60.

Berger, James O., Lawrence D. Brown, and Robert L. Wolpert. 1994. "A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing." *Annals of Statistics* 22:1787–1807.

Berry, Scott M., Bradley P. Carlin, and Jason Connor. 2010. "Bias and Trials Stopped Early for Benefit." *Journal of the American Medical Association* 304:156.

Cox, David. 1958. "Some Problems Connected with Statistical Inference." *Annals of Mathematical Statistics* 29:357–72.

Cumming, Geoff, and Sue Finch. 2005. "Inference by Eye: Confidence Intervals and How to Read Pictures of Data." *American Psychologist* 60:170–80.

Food and Drug Administration. 2010. "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials." http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm.

Goodman, Steven N. 1999. "Toward Evidence-Based Medical Statistics." Pt. 1, "The P Value Fallacy." *Annals of Internal Medicine* 130:995.

———. 2007. "Stopping at Nothing? Some Dilemmas of Data Monitoring in Clinical Trials." *Annals of Internal Medicine* 146:882.

Goodman, Steven N., Donald Berry, and Janet Wittes. 2010. "Bias and Trials Stopped Early for Benefit." *Journal of the American Medical Association* 304:157.

Kiefer, Jack. 1977. "Conditional Confidence Statements and Confidence Estimators." *Journal of the American Statistical Association* 72:789–808.

Lau, Wan-Yee, et al. 1999. "Adjuvant intra-Arterial Lipiodol-Iodine-131 for Resectable Hepatocellular Carcinoma: A Prospective Randomised Trial." *Lancet* 353:797–801.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G., and Michael Kruse. 2001. "Principles of Inference and Their Consequences." In *Foundations of Bayesianism*, ed. David Corfield and Jon Williamson, 381–421. Dordrecht: Kluwer Academic.

Montori, Victor M., et al. 2005. "Randomized Trials Stopped Early for Benefit: A Systematic Review." *Journal of the American Medical Association* 294:2203–9.

Moyé, Lemuel A. 2008. "Bayesians in Clinical Trials: Asleep at the Switch." *Statistics in Medicine* 27:469–82.

Mueller, Paul S., et al. 2007. "Ethical Issues in Stopping Randomized Trials Early because of Apparent Benefit." *Annals of Internal Medicine* 146:878–81.

Nardini, Cecilia. 2013. "Monitoring Clinical Trials: Benefit or Bias?" *Theoretical Medicine and Bioethics* 34:259–74.

Pocock, Stuart, and Ian White. 1999. "Trials Stopped Early: Too Good to Be True?" *Lancet* 353:943–44.

Royall, Richard. 1997. *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Seidenfeld, Teddy. 1981. "On After-Trial Properties of Best Neyman-Pearson Confidence Intervals." *Philosophy of Science* 48:281–91.

Sprenger, Jan. 2009. "Evidence and Experimental Design in Sequential Trials." *Philosophy of Science* 76:637–49.

Wald, Abraham. 1947. *Sequential Analysis*. New York: Wiley.

Worrall, John. 2007. "Evidence in Medicine and Evidence-Based Medicine." *Philosophy Compass* 2:981–1022.

———. 2008. "Evidence and Ethics in Medicine." *Perspectives in Biology and Medicine* 51:418–31.