Bias and Conditioning in Sequential Medical Trials

Cecilia Nardini* Jan Sprenger†

Abstract

Randomized Controlled Trials (RCTs) are currently the gold standard within evidence-based medicine. Usually, they are conducted as *sequential trials* allowing for monitoring for early signs of effectiveness or harm. However, evidence from early stopped trials is often charged with being biased towards implausibly large effects (e.g., Bassler et al. 2010). 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 a shift from unconditional hypothesis tests in the style of Neyman and Pearson to *conditional hypothesis tests* (Berger, Brown and Wolpert 1994) gives a superior appreciation of the obtained evidence and significantly improves 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.

Monitoring refers to the analysis of data in sequential trials carried out as they accumulate, open to the possibility of stopping the trial before the planned conclusion. By terminating a trial when overwhelming evidence for the effectiveness or harmfulness of a new drug is available we can bound the prohibitive costs of a medical trial and protect in-trial patients

*University of Milan and European Institute of Oncology (IEO), Campus IFOM-IEO, Via Adamello, 16, 20139 Milan, Italy. Email: cecilia.nardini@ieo.eu

†Tilburg Center for Logic and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl

against receiving inferior treatments. Thus, monitoring contributes to meeting ethical and epistemic requirements that clinical investigators are confronted with.

However, monitoring in sequential trials also gives rise to a number of fascinating methodological debates. First, 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. Second, the early termination of sequential trials raises a bulk of concerns: For instance, is it ethically mandatory to stop a trial that indicates the possibility of serious adverse effects, jeopardizing the health of actual patients? Or should the treatment be continued in order to avoid that a successful drug is prematurely rejected, which would deprive future patients of an effective cure?

While we cannot adjudicate these far-reaching questions, we follow 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". Thus, we will analyze the statistical methodology of sequential medical trials, focussing on evidence provided by trials *stopped early for benefit*. In the medical literature, such evidence often meets skeptical reactions:

> RCTs stopped early for benefit [...] show implausibly large treatment effects, particularly when the number of events is small. These findings suggest 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 with "appreciable overestimates of effect". However, we do not share the pessimistic conclusion of these authors. While we believe that some of their criticisms of experimental practice in medicine are valid, we believe that the main issue is not a bias inherent in stopping early for benefit, but the 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.

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 (Sect. 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 (Sect. 3). Then we show that *conditional* *frequentist tests* (e.g., Berger, Brown and Wolpert 1994; Berger 2003) reconcile the need for valid post-experimental appraisal of the evidence with preference for frequentist methods and performance measures in the regulatory framework of medical trials (Sect. 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 (Sect. 5).

2 Stopping on a random high?

The practice of stopping RCTs early for benefit has been subject to severe epistemological criticism: trials stopped early for benefit 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. The more recent study by Bassler et al. (2010) shows that truncated trials report significantly higher effects than trials that were not stopped early.

The danger in stopping a trial for apparent benefit consists in promoting a treatment that is actually less efficacious. For instance, 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 practice, the problem with truncated RCTs is often aggravated by improper reporting: crucial elements of trial design such as sample size, points of the interim analysis, or possible ex-post adjustments of effect estimates are missing in a majority of published trials (Montori et al. 2005).

The aforementioned objections severely threaten the reliability of early stopped trials, as well as their reputation in the medical community. Thus, if investigators wish to stop a trial early, they might do so at the risk of ending up with a result that the medical community does not trust. This situation threatens to nullify the possible advantages of monitoring mentioned in the introduction.

The claim of bias made against trials that stop early is based upon an argument that is known in the medical literature under the name of "stopping on a random high". The argument builds on the consideration that evidence suggestive of a strong treatment effect can be observed just by chance. Thus, if several interim analyses are performed, sometimes the trial will be stopped for benefit just by chance, exaggerating a small or null effect. It may even be the case that the trend would vanish or even reverse, if the trial were continued, as happened in the leukemia example mentioned earlier.

The validity of this argument has been questioned by several methodologists, especially by those that are familiar with a Bayesian framework. Goodman, Berry and Wittes (2010) argue that the difference observed in the metastudies of Montori et al. (2005) and Bassler et al. (2010) was 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 early stopped 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.

Is there a methodologically sound way to account for the worry expressed by the "stopping on random high" intuition? We think that the uneasiness in the medical community is not so much about stopping early, but about trials with implausibly large effects – these effects require, in the words of Mueller et al. (2007), "astute clinicians" to make an appropriate interpretation of the results. In the upcoming section we will argue that this uneasiness is caused by the Achilles' heel of statistical methodology in sequential medical trials: the subscription to unconditional inference procedures.

3 Problems with unconditional inference in sequential medical trials

Sequential medical trials usually control the reliability of a testing procedure from a pre-experimental 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, frequentist statisticians and philosophers of science have argued that if the sampling plan is violated, the error probabilities cannot be properly controlled and are actually inflated far beyond acceptable (Mayo and Kruse, 2001).

However, adherence to a proper sequential sampling plan is not sufficient to secure a reliable result. As mentioned at the end of the last section, prior knowledge or empiricallybased prior expectations are highly relevant for sound decision-making in the medical arena (cf. Mueller et al. 2007). Yet, at the present state they do not enter the decisions that are ultimately made, except in a methodologically unsatisfactory *ad hoc* way.

In this respect, Bayesian methods have the potential to alleviate the problems with monitoring discussed above. Bayesian reasoners assign a *prior probability distribution* over the values of the parameter of interest (e.g., relative risk reduction). This distribution represents their subjective uncertainty about the true value of the parameter. By means of Bayes' Theorem, this distribution is updated to a *posterior distribution* that synthesizes the observed evidence with the background knowledge.

Goodman (2007) argues that the inclusion of relevant prior information inherent in the Bayesian framework provides a natural way to account for the relevance of contextual knowledge in medical decision-making. From a Bayesian point of view, successful previous studies on a treatment make a positive result for the current trial more expected and thus support the decision to stop early, while on the other hand, negative results of other studies throw a skeptical light on significant observed effects. Thus, unexpected results will be balanced by the prior and lead to a more conservative conclusion than if Bayesian methods had not been used.

In particular, it can be explained that truncated trials provide, ceteris paribus, less *confidence* than trials with a comparable effect size that were completed. 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 – are naturally accounted for.

Despite the advantages just outlined, there are some serious counterarguments to the viability of Bayesianism in clinical trials. A first issue is that some of the philosophical implications of Bayesian inference – such as the evidential, post-experimental irrelevance of experimental design – conflict with the need to carefully plan and conduct sequential medical trials. 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 (cf. FDA 2010).

Moyé (2008) has also highlighted a non-sociological point: the problematic specification of a prior belief function.[1] While "objective", non-informative priors (Jeffreys 1961; Bernardo

[1] Similar worries arise regarding the definition of an appropriate loss function required for a Bayesian decision

1979) respond too easily to implausibly large effect sizes, the history of medical trials shows that subjective beliefs about the efficacy of a drug are all too often overturned by surprising findings. The latter problem hampers the use of properly subjective priors and, according to Moyé, it persists even if data from meta-analyses are taken into account.

We consider these worries legitimate and we think they may represent a crucial counter-indication to the use of Bayesian methods in healthcare assessment, even though some of the issues are regulatory rather than epistemological. That said, we believe that the often-cited antagonism between Bayesians and frequentists rests on the false presumption that either of the two is right while the other is wrong. In fact, we suggest to replace that antagonism by the contrast between *conditional* and *unconditional* procedures. By "conditional", we refer to statistical procedures that quantify the conclusiveness of a test result by conditioning on part of the observed data, while "unconditional" refers to the absence of such conditioning.

Arguably, what is most disturbing to the medical community is the fact that, according to current unconditional 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. By contrast, in a conditional perspective, the error associated with a particular conclusion depends on the observed data: the larger the observed difference is, the lower the probability that the null is rejected erroneously.

Practitioners that rely on unconditional inference have an hard time to find informative and reliable *post-data 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, "confidence interval" is a misnomer: 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

model.

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, it should not come as a surprise that some confidence intervals include the entire sample space, raising the question of what we have actually learned (cf. Seidenfeld 1981).

Finally, we contend that the *unconditional* nature of Neyman-Pearson hypothesis tests is the culprit for their shortcomings. To motivate and to defend 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 i.i.d. 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 exmperiment. 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%).

Neyman-Pearson methodologists may be inclined to dismiss the second test because not all of its components are tests at the 5% level. However, from an unconditionalist (pre-experimental) viewpoint, only the overall error rates should count. Here, the superior power features speak for the second, gerrymandered test. This problem reveals the tension between the pre-experimental 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, pre-experimental error probabilities can qualify the goodness of an inference. In the next section we will see how conditioning on the relevant chunks of information overcomes the problems of unconditional inference and resolves the methodological confusion about interpreting truncated RCTs, without altering or 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 co-authors.

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 (that is, 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.

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 Bayesian 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 a conditional error probability. 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 F_0 and F_1 be the corresponding cumulative distribution functions.

$$F_0(x) := P_{H_0}(X \leq x) \qquad\qquad F_1(x) := P_{H_1}(X \leq x)$$

Let the *Bayes factor* $B(x) := f_0(x)/f_1(x)$ be the ratio of the probability density functions, and let

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

It is easy to check that $\mathcal{X}_s$ has the same probability density under H_0 and H_1, for all values of s. The intuitive idea is that any $\mathcal{X}_s$ contains two values that have the same strength of evidence under H_0 and H_1. The outcome space is thus partitioned into subsets $\mathcal{X}_s$.

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

$$T^*(X) = \begin{cases} \text{Reject } H_0 & \text{if } B(X) < 1 \\ \text{Accept } H_0 & \text{if } B(X) \geq 1 \end{cases}$$

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

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

where the latter equalities have been proven by Berger, Brown and Wolpert (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.

Before moving to the Bayesian interpretation of conditional tests, we would like to briefly discuss 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 it is of course possible to change the rejection region according to contextual requirements.

Second, the use of the Bayes factor may indicate that the conditional test is actually a Bayesian test in frequentist cloths. However, $B(X)$ possesses a frequentist interpretation, too, since it identifies the most powerful frequentist test in the simple vs. simple testing problem.[2]

Third, 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, Boukai and Wang (1997) have extended conditional tests to simple vs. composite testing problems, and in particular, to the two-sided null hypothesis testing problems that frequently occur in RCTs.

We now explain why T^* is also a valid Bayesian test. Assume that the prior probabilities are balanced: $P(H_0) = P(H_1) = 1/2$. This may be defended as a useful neutrality assumption. Then, the posterior probability of H_0 and H_1 can be written as

$$\begin{aligned} P(H_0|x) &= \left(1 + B(x)^{-1}\right)^{-1} = \frac{B(x)}{1 + B(x)} \\ P(H_1|x) &= (1 + B(x))^{-1} = \frac{1}{1 + B(x)} \end{aligned}$$

Thus, we see that the posterior probabilities of H_0 and H_1 correspond to the conditional error probabilities for rejecting H_0 and H_1, respectively. Indeed, the decision to accept H_0 will be wrong whenever H_1 is actually true, that is, with probability $1/1 + B(x)$. Thus, Bayesians and frequentists can conduct the same (conditional) test and obtain the same numerical conclusions. But for the purposed of medical *practice*, philosophical questions about

[2]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 post-data error probabilities as T^*.

the interpretation of probability are clearly secondary as long as there is methodological agreement on procedures and post-experimental data assessment (cf. Berger 2003). In this sense, conditional inference is a genuine reconciliation of Bayesian and frequentist methodology and a real asset for practitioners.

As a last point in indicating the advantages provided by the conditional frequentist framework, we discuss its application to sequential analysis. The proponents of conditional testing have stressed repeatedly that one of the main motivations of conditional inference was the desire to improve upon the practice of sequential testing, particularly in medicine. Here the benchmark is Wald's (1947) famous Sequential Probability Ratio Test, that is

$$T^N(X): \begin{cases} \text{Reject } H_0 \text{ and stop sampling} & \text{if } B(X_1, \ldots, X_N) \leq C^- \\ \text{Accept } H_0 \text{ and stop sampling} & \text{if } B(X_1, \ldots, X_N) \geq C^+ \end{cases}$$

with associated (unconditional) error probabilities

$$\alpha = P_{H_0}(B(X_1, \ldots, X_N) \leq C^-)$$
$$\beta = P_{H_1}(B(X_1, \ldots, X_N) \geq C^+).$$

While these unconditional error probabilities are (i) misleading and (ii) very hard to calculate, Berger, Brown and Wolpert (1994) have suggested a conditional interpretation of this test, choosing C^+ and C^- such that $F_0(C^-) = 1 - F_1(C^+)$, and reporting conditional error probabilities

$$\alpha(B(X_1, \ldots, X_N)) = \frac{B(X_1, \ldots, X_N)}{1 + B(X_1, \ldots, X_N)} \tag{2}$$
$$\beta(B(X_1, \ldots, X_N)) = \frac{1}{1 + B(X_1, \ldots, X_N)}. \tag{3}$$

Thus, the conditional framework can be straightforwardly applied to sequential medical trials, and it has significant advantages. 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. It seems reasonable to maintain that 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.

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 a 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 should not be misunderstood as the claim that pre-data analysis and experimental design are superfluous. Unfortunately, Berger, Brown and Wolpert (1994: 1803) make a claim into that direction, but given the strong emphasis on careful design by methodologists and regulatory bodies (cf. Moyé 2008; FDA 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.

There are also interesting implications for the philosophy of statistics: if the "error statisticians" (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 Sect. 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 paper.

5 Conclusions

In this paper 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 towards 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.

This approach holds considerable promise for the interpretation of early stopped trials in medicine. The possibility of post-data 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 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

[1] Bassler, D., Briel, M., Montori, V., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansdell, D., Walter, S., Guyatt, G., N Flynn, D., et al.: Stopping randomized trials early for benefit and estimation of treatment effects. JAMA **303**(12), 1180–1187 (2010)

[2] Berger, J.: Could Fisher, Jeffreys and Neyman have agreed on testing? Statistical Science **18**(1), 1–12 (2003)

[3] Berger, J., Boukai, B., Wang, Y.: Unified frequentist and Bayesian testing of a precise hypothesis. Statistical Science **12**(3), 133–160 (1997)

[4] Berger, J., Brown, L., Wolpert, R.: A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. The Annals of Statistics **22**(4), 1787–1807 (1994)

[5] Bernardo, J.: Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society. Series B (Methodological) pp. 113–147 (1979)

[6] Berry, S., Carlin, B., Connor, J.: Bias and trials stopped early for benefit. JAMA **304**(2), 156 (2010)

[7] Cox, D.: Some Problems Connected with Statistical Inference. Annals of Mathematical Statistics **29**(2), 357–372 (1958)

[8] Cumming, G., Finch, S.: Inference by eye: Confidence intervals and how to read pictures of data. American Psychologist **60**(2), 170 (2005)

[9] Goodman, S.: Toward evidence-based medical statistics. 1: The P value fallacy. Annals of Internal Medicine **130**(12), 995 (1999)

[10] Goodman, S.: Stopping at nothing? Some dilemmas of data monitoring in clinical trials. Annals of Internal Medicine **146**(12), 882 (2007)

[11] Goodman, S., Berry, D., Wittes, J.: Bias and trials stopped early for benefit. JAMA **304**(2), 157 (2010)

[12] Jeffreys, H.: Theory of Probability. Clarendon Press, Oxford (1961)

[13] Kiefer, J.: Conditional confidence statements and confidence estimators. Journal of the American Statistical Association pp. 789–808 (1977)

[14] Mayo, D.: Error and the growth of experimental knowledge. University of Chicago Press (1996)

[15] Mayo, D., Kruse, M.: Principles of Inference and their Consequences. In: Foundations of Bayesianism. Kluwer Academic Publishers, Netherlands (2001)

[16] Montori, V., Devereaux, P., Adhikari, N., Burns, K., Eggert, C., Briel, M., Lacchetti, C., Leung, T., Darling, E., Bryant, D., et al.: Randomized trials stopped early for benefit: A systematic review. JAMA **294**(17), 2203 (2005)

[17] Moyé, L.A.: Bayesians in clinical trials: Asleep at the switch. Statistics in Medicine **27**, 469–482 (2008)

[18] Mueller, P., Montori, V., Bassler, D., Koenig, B., Guyatt, G.: Ethical issues in stopping randomized trials early because of apparent benefit. Annals of Internal Medicine **146**(12), 878 (2007)

[19] Royall, R.: Statistical Evidence: A Likelihood Paradigm. Chapman & Hall, London (1997)

[20] Seidenfeld, T.: On after-trial properties of best Neyman-Pearson confidence intervals. Philosophy of Science **48**(2), 281–291 (1981)

[21] US Food and Drug Administration: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (2010). Available at `http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm`. Last access 26/01/2012

[22] Wald, A.: Sequential analysis. Wiley, New York (1947)

[23] Worrall, J.: Evidence in Medicine and Evidence-Based Medicine. Philosophy Compass **2**(6), 981–1022 (2007)

[24] Worrall, J.: Evidence and Ethics in Medicine. Perspectives in Biology and Medicine **51**(3), 418–431 (2008)