Stopping Rules and Data Monitoring in Clinical Trials

Roger Stanev
Department of Philosophy
University of British Columbia

Abstract

Philosophers subscribing to particular principles of statistical inference need to be aware of the limitations and practical consequences of the statistical approach they endorse. The framework here proposed, together with methodological guidelines, allows disparate statistical approaches to emerge in their appropriate context while providing important considerations for deciding on trial conduct. While these considerations do not amount to stopping rules, they would assist data monitoring committees in judging their position with regard to necessary precautionary interpretation of interim data. My conclusion raises suspicions about philosophies of science that promote a universal principle of statistical inference applied to clinical trials.

Introduction

Stopping rules—rules dictating when to stop accumulating data and start analyzing it for the purposes inferring from the experiment—divide Bayesians, Likelihoodists and classical statistical approaches to inference. Although the relationship between Bayesian philosophy of science and stopping rules can be complex (cf. Steel 2003), in general, Bayesians regard stopping rules as irrelevant to what statistical inference should be drawn from the data. This position clashes with classical statistical accounts. For orthodox statistics, stopping rules do matter to what inference should be drawn from the data. "The dispute over stopping rule is far from being a marginal quibble, but is instead a striking illustration of the divergence of fundamental aims and standards separating Bayesians and advocates of orthodox statistical methods." (Steel 2004, 195)

But philosophers who subscribe, on theoretical grounds, to particular principles of statistical inference need to recognize the limitations of the statistical approach they endorse when it comes to important matters, such as the conduct of randomized clinical trials (RCTs). In broadest terms, I am concerned with the following problem: what if no single statistical approach is best-suited to address all the necessary demands of clinical research? The paper focus on a specific version of this problem: the apparent inability of existing statistical approaches to accommodate two such demands. The first is that clinical trials incorporate some basic stopping rule, and the second is that clinical trials incorporate policies for early termination (at times in violation of the basic stopping rule). While many statistical approaches can meet one of these demands, no extant approach appears capable of meeting both. I suggest that this type of

predicament requires new ways of thinking about the problem in order to give credit to distinct approaches where it might be due. Rather than solving the problem by formulating yet another universal paradigm for statistical inference, this paper proposes a framework together with methodological guidelines that provide important considerations for deciding on trial conduct.

The paper proceeds as follows. Section 2 introduces the problem of stopping rules and the problem of early stopping of RCT. Data monitoring procedure is then presented as a means of addressing some of the problems in early stopping RCT. As an example of such trials, Section 3 depicts the monitoring experience of a data monitoring committee deciding to early stop an RCT based on the unexpected low event rate observed during interim analysis. Section 4 introduces the decision framework and how the framework treats the monitoring of RCTs. Section 5 discusses some simulations through the framework and how it can assist in comparing monitoring decisions occurred in practice. Section 6 concludes.

Stopping rules and the monitoring of RCTs

In a nutshell, the difference among the philosophies of statistics with regards to the importance of stopping rules is that these rules do not impact likelihoods for Bayesians,[1] but they impact a procedure's error-probabilities for statistical accounts of inference, such as Mayo's (1996) error-statistics (ES). A common Bayesian response to the irrelevance of stopping rules is that these rules reflect private intentions of the experimenter. These intentions—which we presumably have no access to—should not matter for an account of statistical evidence. For scientists—ideal scientists—would not regard such personal intentions as proper influences on the support which data *x* give to a hypothesis H.

In contrast, Mayo and Kruse (2001) for instance, argue that to the error statistician the situation is the reverse of what we find with the likelihood principle. "[T]he stopping rule is relevant because the persistent experimenter is more likely to find data in favor of H, even if H is false, than one who fixed the sample size in advance." (2001, 389) For ES, stopping rules as test specifications cannot be relegated to a particular experimenter's intention. They are part of the experimental design considerations impacting error-probabilities, which are operating characteristics of the test procedure. A try-and-try-again procedure with an optional stopping point could lead to high or maximal overall significance levels. Stopping rules affect both the reported error rates and what should be expected to happen in subsequent repetitions of the experiment. They are not just important for correctly reporting what occurred in obtaining the

[1] The likelihood principle says that if $P(x|\theta)=cP(x'|\theta)$, where *c* is some positive constant and x and x' similar data from different experiments testing the same hypotheses H about θ, then both data have identical evidential import. E.g. see Mayo (1996).

experimental result, but also in determining what should be expected to occur in subsequent repetitions of the experiment performed. Other experimenters seeking to check or repeat the results observed run the risk of being misled when stopping rules are ignored.

For RCTs, however, the issue over stopping rules is more complex. What if an unanticipated beneficial effect emerges as early as one year into a five year trial? Should we discontinue the trial? If the trial is to be stopped quite early, should the evidence for efficacy be overwhelming? If the trial is stopped halfway through in favor of one treatment, how likely is it that, if the trial were allowed to continue there could be a reverse in trend before the study ends? These are non-trivial matters that data monitoring committees (DMC) often face. Data monitoring is repeated examination of the data as it accumulates, with an eye to possible early termination of the trial.

Differently than stopping rules, *early stopping policies or principles* do (and at times should) override stopping rules. Stopping principles allow the trial to halt in one of three cases: early stop due to harm, efficacy or futility. In RCTs, it is the monitoring rule which is operational, not the stopping rule. Since stopping rules are expressed in non-specific terms (e.g. stop when alpha is less than 0.05 at the first interim for efficacy) the monitoring rule implements the actual stopping rule using a statistical monitoring plan (e.g. alpha spending function), a statistical method (e.g. conditional power) and an ethical norm such as "a favorable balance benefits and harms should exist". The upshot is that while ES has an easier time explaining stopping rules, a decision-theoretic approach may have an edge in accounting for early stopping.

For instance, in most treatment RCTs, in contradistinction to prevention RCTs, subjects suffer from a certain condition seeking alleviation of its consequences. Monitoring the amount of statistical evidence is therefore confined to certain aspects of such consequences, e.g., time to death from the particular disease. Since there is substantial information about mortality (or morbidity) that accrues within the first few years of the trial the focus on evidence is on treatment effect occurring over a relatively short period of time. This means the evidence (e.g., mortality rates dropping) is judged in ways that are sufficient to offset the chances of later finding harmful effects. In this type of RCT, the suggestion seems to be that even if such harmful effects would occur, the DMC would tolerate harmful risks for the sake of evidence found for early benefit. If this characterization of treatment RCT is correct, this indicates that an implicit utility specification is at play.

But even though there is almost unanimity among researchers about the ethical necessity of monitoring interim data, "there are still widely disparate views on early stopping criteria." (Ellenberg 2003, 586) Some—mostly in Europe, says Ellenberg—hold the view that clinical trials should rarely, if ever, stop early since extreme evidence is needed to have an impact on clinical practice. In contrast, others—mostly in the United States—believe that "once it has become clear that the question addressed by the trial has been answered with pre-specified statistical precision, the trial should be terminated." (2003, 586) According to British researcher Stuart Pocock, "in HIV trials, especially in the United States, the push towards individual ethics at the expense of collective ethics" as in, "stopping trials too soon" "has been detrimental to determining the most effective therapeutic policies." (1993, 1466)

Moreover, despite challenges of evidential interpretation that can arise when stopping a trial early, according to a recent systematic review in *JAMA*, the number of RCTs stopped early for benefit has more than doubled since 1990, often failing to adequately report relevant information about the decision to stop, and with clustering of publication occurring mostly in top medical journals (Montori et al, 2005). The problem is no different when it comes to harm. A survey of trials stopped early for harm in HIV/AIDS says that the reporting of methods informing the decision to stop "is deficient in a variety of ways, including lack of stopping guidelines" thus recommending caution when interpreting such trials. (Mills et al, 2006)

Therefore, in addition to stopping rules, early termination furnishes a second important test case for competing statistical approaches. Should the statistical analysis be adjusted for the fact that interim data checks have been performed in the past or that future reviews might be undertaken? It seems that Bayesians and error-statisticians might have very different answers to this question, and they are not the same answers as for the issue of stopping rules. The next section develops this contrast via a case of early stopping for futility based on conditional power.

Conditional power and the early stopping for futility

The statistical method of conditional power can be used to assess whether an early unfavorable trend can still reverse itself, in a manner that is sufficiently enough to show a statistically significant favorable trend at the end of the trial. It allows one to assess how likely the early trend might reverse, which could be informative for scenarios involving early unfavorable trend, early beneficial trend, or whether the trial should be extended beyond its planned termination.

The frequentist version of the method computes the probability of rejecting the null-hypothesis under a *pre-specified* effect size (θ) conditional on the data observed up to that moment. DMCs have relied on such approach, not in an exclusive manner, but in a complementary and important way in a variety of RCTs when making informed decisions on whether to continue or stop trials, whether for futility, harm or efficacy.

The multi-center RCT conducted in 1994 evaluating an intervention that could prevent a brain-infection in HIV+ individuals is a good example of how conditional power was important in assisting the DMC. Because *toxoplasmic encephalitis* (TE) is a major cause of morbidity and mortality among patients with HIV/AIDS and it had already been known by then to be "the most common cause of intra-cerebral mass lesions" (Jacobson et al 1994, 384) an RCT was set to evaluate pyrimethamine in the prevention of TE.

The study was designed with a target of 600 patients followed for a period of 2+1/2 years, with an estimate that "a 50% reduction of TE with pyrimethamine could be detected with

a power of .80 at a two-sided significance level of .05 if 30% of patients given placebo developed TE during follow-up." (Jacobson et al 1994, 385) Survival was the primary end point of the study. By March 1992, at the time of the forth interim analysis, while patients were still being enrolled, the investigators found that:

> "[T]he committee recommended that the study be terminated because the rate of TE was much lower than expected in the placebo group and was unlikely to increase appreciably during the planned duration of the study and because there was a trend toward rates of both TE and death being higher for patients given pyrimethamine than for patients given placebo. Thus, it was thought unlikely that pyrimethamine as used in this study was an effective prophylaxis against TE." (Jacobsen et al 1994, 386)

Even though the original publication did not report on the type of conditional power computations that led to the early termination decision, nor did it specify the statistical monitoring plan adopted, it is clear from the publication that due to the unexpected low TE event rates observed during interim analysis, which compromised the original power of the study, and due to an early unfavorable trend in survival, the DMC recommended the early stopping of the trial. But the decision to early stop was not unanimous and far from uncontroversial.

In a subsequent article authored by the original investigators, they explain that "a Haybitte-Peto interim analysis monitoring plan for early termination" was used by the DMC. (Neaton et al 2006, 321) Despite the DMC recommending the trial to no longer continue "the chair advocated continuing the study due to the uncertainties about the future TE event rate and about the association of pyrimethamine with increased mortality" while "the DMC reaffirmed their recommendation to stop the trial." (Neaton et al 2006, 326) Among the "lessons learned"

investigators expressed a desire "that procedures should be in place for adjudicating differences of opinion about early termination." (Neaton et al 2006, 327)

Based on this epidemiologic study, two important points are in order concerning the dispute between the competing philosophies of statistics. First, from an initial glance at the episode, it seems that from the point of view of ES philosophy of statistics, the episode 'fits' squarely with its philosophy. That is, since the DMC adopted a group sequential boundary method—Haybittle-Peto interim monitoring—which aims at 'spending' alpha (type I erro) in such way as to control the overall type I error due to the multiple interim analysis, the episode seems to count as an instance in favor of the ES philosophy. "Due to the conservatism of this boundary at each interim analysis, the adjustment to the final critical value to obtain an overall one-sided 0.025 significance level is very small." (Ellenberg et al 2002, 125) The concern for the control of error probabilities would seem vindicated here.

On the other hand, the repeated applications of the conditional power method during the ongoing trial could be a problem to the ES philosophy. Because conditional power computation could eliminate prematurely the possibility of detecting an effect of interest, it does increase the probability of false negatives, i.e. of failing to reject the H_0 in favor of the alternative hypothesis. If, however, the conditional power calculation under consideration is smaller than 0.2—relative to the originally hypothesized power—provided the original power was 0.85 or higher, "the increase in the rate of false negative error is negligible." (Ellenberg et al 2002, 130) If this is correct, then for most practical purposes where the study was well powered to detect the alternative hypothesis of interest, the ES worry about having to rely on a method that increases

the probability of false negatives, is somewhat mitigated, although technically a violation of the controlling error probabilities principle.

Second, assuming that the chair and the DMC used the same statistical monitoring plan in arriving at their respective recommendations, including same ethical principle, and similar judgments about the consequences of continuing or stopping the trial, how can we account for the difference between their recommendations, assuming both recommendations were reasonable?

One way is to base the difference on "the uncertainties about the future TE event rate" expressed by the chair of the study. The consideration suggests that the chair might have adopted a different statistical method for computing conditional power when arriving at his recommendation. Whereas the DMC used the frequentist conditional power based on a single pre-specified effect size, the chair computed conditional power considering a range of reasonable effect sizes with different weights to them. This method is a hybrid Bayesian-frequentist method of conditional power, and is known as the *predictive power* approach, which involves averaging the conditional power function over the range of reasonable effect sizes.

It is under these assumptions that we turn our attention to the decision framework. The framework assists us in comparing the performance of different statistical monitoring plans, with respect to a maxim while also assisting in rationalizing the decisions based on them.

The comparative rationale driving the framework is this: the decision to terminate a trial is based on whether ending it has a lower expected loss than continuing, where the expectation is with respect to the interim posterior probability, and on the consequence of continuing considering a fixed set of future actions. By using this rationale we can compare, for instance, given the losses incurred by the mistreatment of patients, how much the DMC recommendation differed from the chair's under a certain maxim.

1. **Decision Theoretic Framework**

The evaluation of statistical monitoring plans is treated as a decision under risk, represented as a 4-tuple (Θ, A, Y, L). Θ is our parameter space, A is the set of actions, Y is the data model, and L is a specific loss function. $\delta(y): Y \rightarrow A$ is a statistical decision monitoring rule.

Keeping the example simple, we treat the event of interest as a binary event. The example involves dichotomous observations, two treatments, two loss functions, two statistical decision rules, one maxim, three types of actions, and easy to compute numbers.

Suppose N individuals suffer from HIV+. There are two treatments: standard treatment T_1 and experimental treatment T_2. In order to find which of the two treatments is more effective we conduct an RCT on $2n$ of the total N patients, with n assigned to each treatment. Suppose the remaining $N - 2n$ patients receive the treatment selected as the more effective of the two when the trial ends, unless no treatment is declared superior, in which case remaining patients will be treated with standard treatment T_1.

We assume statistical monitoring rules that permit termination after $n/2$ have been treated. The RCT has a single interim analysis halfway through the intended trial. We treat the monitoring of 10 pairs of patients (n=10), 5 patients assigned to each treatment by interim analysis, and 10 patients assigned to each treatment by final analysis, with N=100.

4.1 Observation model

For each treatment we observe a single outcome: whether or not the patient recovers. The probability of recovery with T_i is θ_i (i=1,2) (assumed constant from patient to patient). y_1 and y_2

denote the number of recoveries among the *n* patients using T_1 and T_2 respectively. We model our observations so as to follow a binomial distribution:

$$f(y_i \mid \theta_i) = \binom{n}{y_i} \theta_i^{y_i} (1-\theta_i)^{n-y_i} \quad (i = 1, 2)$$

Since we are interested in the difference between T_1 and T_2, let θ stand for the difference in response rates between T_1 and T_2: $\theta = \theta_2 - \theta_1$

Thus, the joint distribution of y_1 and y_2 given θ_1 and θ_2 is:

$$f(y \mid \theta) = \prod_{i=1}^{2} \binom{n}{y_i} \theta_i^{y_i} (1-\theta_i)^{n-y_i} \ (i = 1, 2)$$

y and θ denoting (y_1, y_2) and (θ_1, θ_2), respectively; $y \in Y$ and $\theta \in \Theta$.

4.2 Parameter space Θ

Let us assume that the recovery rate of T_1 (θ_1) is 0.5 and that of T_2 (θ_2) is unknown. Moreover, assume that θ takes one of two possible values:

H_0: $\theta=0$ (T_2 is equal to T_1 i.e., $\theta_2=\theta_1=0.5$), or

H_1: $\theta=0.2$ (T_2 is more effective than T_1 i.e., $\theta_2=0.7$)

4.3 Set A of possible actions

At the interim analysis, we can act one of three ways:

a_1: Stop and declare "T_2 is more effective than T_1",

a_2: Stop and declare "T_2 is equal to T_1",

a_3: Continue the trial.

At the end of the trial, we can act one of two ways:

a_1.f: Stop and declare "T_2 is more effective than T_1",

a_2.f: Stop and declare "T_2 is equal to T_1".

At either the interim analysis or the end of the trial, the choice of action among the alternatives is made on the basis of sample data y_1 and y_2. Specific sample data expressed by the pair (y_1, y_2) leads to the choice of action to take, a choice that depends ultimately on the decision monitoring rule.

A decision monitoring rule is a function from sample data y into the set A of allowed actions $\{a_1, a_2, a_3, a_1.f, a_2.f\}$. They specify how actions are chosen, given observation(s) y.

4.4 Decision monitoring rule $\delta(y): Y \rightarrow A$

We first consider the DMC decision monitoring rule. This rule is based on frequentist properties by controlling error probabilities. We use an instance of a group sequential monitoring procedure much like the one adopted by the DMC. This rule aims at controlling type I error by having it to be no more than 0.05 (approx.). For this, DMC considers rejecting H_0 (when H_0 is true) for values $(y_2 - y_1) \geq 4$,[2] otherwise they do not reject H_0.

For the study chair decision monitoring rule, we shall assume they used a Bayesian rule. For instance, if, the chair wanted to compute conditional power considering a range of reasonable effect sizes with different weights to them, then a hybrid Bayesian-frequentist method

[2] $P((y_2 - y_1) \geq 4|H_0)=0.058$, computed according to the joint probability distribution. E.g., $P((y_2=7, y_1=3)|H_0)=(0.117)^2=0.014$.

of conditional power can appropriately be called for. With predictive power, the monitoring rule requires the specification of prior information about θ expressed in terms of a prior probability mass function $p(\theta)$. For simplicity, we consider neither an optimistic nor a pessimistic set of priors, but 'flat' priors: $p(\theta=0)=p(\theta=0.2)=0.5$ When using a Bayesian monitoring rule, the evidence provided by data y is contained in the likelihood ratio which is multiplied by a factor, the ratio of prior probabilities, producing the ratio of posterior probabilities. Therefore, when choosing between H_0 or H_1 on the basis of y, the rule chooses the hypothesis with larger posterior probability. For instance, putting it in terms of rejecting H_0, the rule may reject H_0 when the likelihood ratio is less than 1. If so, with flat priors, the rule rejects H_0 as long as $y_2 \geq 6$; otherwise it does not reject H_0.

In order to make both monitoring rules comparable, despite different philosophies, we choose a cutoff point during the interim analysis—which is our point of contention—for the chair's rule to be as close as possible to the DMC's one, while keeping with our choice of easy to compute numbers.

We start with the following loss function: the 'ethical' loss function, $L_E(\theta,\mathbf{a})$. From its perspective, the goal is to compare both treatments by *paying a penalty for each patient assigned with inferior treatment*. One unit is counted for each patient assigned the inferior treatment.

One way to think of $L_E(\theta,\mathbf{a})$ is to see it with respect to a particular patient. If H_0 is true, whether the patient is given T_1 or T_2 the loss incurred with such treatment is 0 (since the treatments are equivalent), but when H_1 is true, if the patient is given the inferior treatment T_1 the

loss is ***d*** (which we assume for now as 1 unit) otherwise, if given T_2 (superior treatment) the loss is 0.

We also make $L_E(\theta,\mathbf{a})$ sensitive to the "effect size" $|\theta_2-\theta_1|$. By "effect size" we mean the percentage difference of recoveries between treatments. Thus, assuming H_1 is true ($\theta=0.2$), for every n=5 patients (every segment of interim analysis), the single loss unit is the loss incurred (associated) from 1 fewer patient having a positive recovery, the result that researchers should have obtained (or could have expected) had researchers continued with the trial.

4.6 Maxim

Average loss is obtained by averaging the loss function over all possible observations:

$$E_y[L(\theta,\delta(y))\,|\,\theta] = \sum_{y_i \in Y} L(\theta,\delta(y)) f_y(y_i\,|\,\theta)$$

If we have prior information about θ which can be expressed in terms of a prior probability mass function $p(\theta)$ the Bayes' risk of decision rule $\delta(y)$ is the expectation of the average loss over possible values of θ:

$$r(\delta(y)) = \sum_{\theta\in\Theta}\sum_{y\in Y} L(\theta,\delta(y)) f_y(y\,|\,\theta)\, p(\theta)$$

2. Discussion

The decision to stop or continue is made by weighing the consequences of possible actions, averaging over the distribution of future observations—as yet unobserved results have probability distribution. Consider the situation of the DMC at the end of their first interim analysis, with 0.05 p-value in favor of T_2 using their conditional power rule. How does it compare with the chair's decision monitoring procedure given our ethical loss function?

Figure 1 show for a set of different effect sizes, the weighted-losses [3] for stopping and continuing, according to DMC and the chair's monitoring procedures, when H_1 is true.

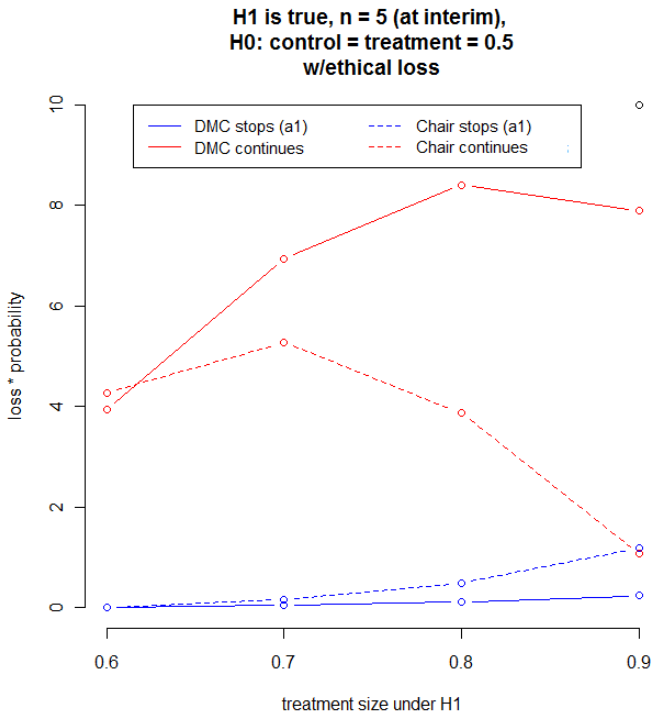

Figure 1: Weighted-losses of stopping (acting a1, in blue) and continuing (averaging actions a1.f and a2.f, in red) for each procedure, when H_1 is true.

[3] Weighted-loss=(loss)*(probability of taking that action).

If we assume that the DMC decision to stop was based on whether ending it had a lower expected loss than continuing it, where the expectation is with respect to the weighted-losses of continuing considering our fixed set of future actions (either a1.f or a2.f) then, by averaging the weighted-losses of a1.f and a2.f, we can compare the weighted-losses of stopping vs. continuing for a whole range of effect sizes. Notice that for every considered effect size, the DMC decision to stop the trial has a lower expected loss than continuing. We say the decision to stop is undivided. This, however, is not true with the chair's monitoring rule. Given our loss ethical function, and assuming that chair's decision to stop was also based on whether ending it had a lower expected loss than continuing, then, the only effect size that could have warranted the decision to continue was if the treatment size was expected to be high, θ_2=0.9, otherwise the trial should have stopped, since stopping at interim had lower expected loss than continuing.

If, however, we assume that both DMC and the chair decisions were reasonable, and we seek justifying both decisions as being 'equally' ethical, under the very same maxim, then we must consider different loss functions. One contender is the 'scientific' loss function, based on the idea of finding the true state of nature, whatever its consequences. From its point of view, correctly declaring that "T_2 is more effective than T_1" has the same utility to correctly declaring that "T_2 is equal to T_1", even though the two states of nature have quite distinct consequences to present and future patients. Turning it in terms of losses, we may assume *the same penalty for any incorrect conclusion*. Suppose a 10 unit loss.

Now, we can compare the DMC vs. the chair's rule, under the same maxim, but varying losses. **Figure 2** shows how the two monitoring rules perform under Bayes' risk, for the range of effect sizes, under each loss function. By assuming an even set of priors for every pair of hypotheses (each pair having a different effect size), and adopting the same 'ethical' loss function, the DMC monitoring rule outperforms the chair's rule with respect to the Bayes' risk when the effect size is small, θ_2=0.6. With respect to this maxim, the DMC rule is therefore 'more' ethical then the chair's when the effect size is small. But, for all other effect sizes, the Bayes' risk for the chair's rule is smaller than the DMC rule. No surprise given the fact that the chair's rule is a Bayesian monitoring rule. Obviously, different set of priors would produce different Bayes' risks. Had the chair been optimistic regarding pyrimethamine's benefits, (e.g. $P(H_1)$=0.8, $P(H_0)$=0.2) chair's Bayes risk curve would have resulted in a shift upwards in **Figure 2**. In that case, for all effect sizes except θ_2=0.9 (when the effect is largest), the DMC decision monitoring would have outperformed chair's rule with respect to Bayes' risk.

The situation is reversed by adopting the 'scientific' loss function. Here, the DMC monitoring rule outperforms chair's rule with respect to the Bayes' risk in all effect sizes, except when effect is large, θ_2=0.9. With respect to this function, the DMC rule is therefore 'more' ethical then chair's except when the effect is large.

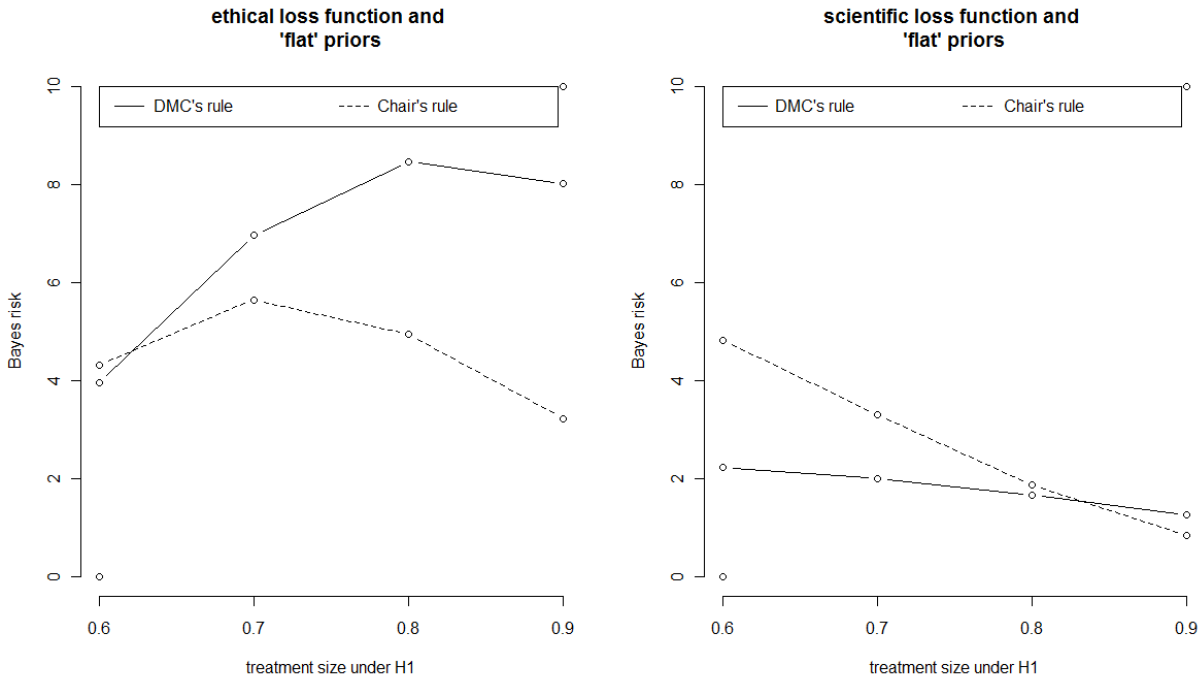

Figure 2 DMC vs. chair's rule under different loss functions.

6. Conclusion

When it comes to RCTs, philosophers subscribing to particular principles of statistical inference should be aware of the limitations and consequences of the methods they endorse. The framework proposed allows disparate monitoring rules to emerge in their appropriate context while providing important considerations for deciding on trial conduct based on a range of factors. The framework leaves us with further questions. Is this an effective avenue for adjudicating between statistical decision rules? Does the framework open some new way of seeing why the DMC and chair recommendations were more than reasonable ones, suggesting perhaps some further social or political motivation behind their recommendation? Do we need another level of ethical decision rule when it comes to adjudicating between statistical monitoring methods? We invite others to contribute.

Bibliography

Ellenberg, S. S., T. H. Fleming, and D. L DeMets (2002), *Data Monitoring Committees in Clinical Trials*, John Wiley&Sons.

Ellenberg, S. (2003), "Are all monitoring boundaries equally ethical?" *Controlled Clinical Trials* **24**:585-588.

Jacobson et al (1994), "Primary Prophylaxis with Pyrimethamine for Toxoplasmic Encephalitis in Patients with Advanced Human Immunodeficiency Virus Disease: Results of a Randomized Trial", *Journal of Infectious Diseases* **169**:384-394.

Neaton, J., Wentworth, D., and Jacobson, M. (2006), "Data Monitoring Experience in the AIDS Toxoplasmic Encephalitis Study", in DeMets, D., Furberg, C., and Friedman, L. (eds.), *Data Monitoring in Clinical Trials*, New York, Springer.

Mayo, D. (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press.

Mayo, D., and Kruse, M. (2001), "Principles of Inference and Their Consequences", in David Corfield and Jon Williamson (eds.), *Foundations of Bayesianism*. Dordrecht:Kuwer Academic Publishers, 381-403.

Montori et al. (2005) "Randomized trials stopped early for benefit: a systematic review" *JAMA*, **294**:2203-09.

Pocock, S.J. (1993), "Statistical and Ethical Issues in Monitoring Clinical Trials", *Statistics in Medicine*, **12**:1459-146.

Steel, D. (2003), "A Bayesian Way to Make Stopping Rules Matter", *Erkenntnis* **58**:213-227.

Steel, D. (2004), "The Facts of the Matter: A Discussion of Norton's Material Theory of Induction", *Philosophy of Science* **72**:188-197.