# Quantitative Framework for Retrospective Assessment of Interim Decisions in Clinical Trials

Roger Stanev, PhD

*This article presents a quantitative way of modeling the interim decisions of clinical trials. While statistical approaches tend to focus on the epistemic aspects of statistical monitoring rules, often overlooking ethical considerations, ethical approaches tend to neglect the key epistemic dimension. The proposal is a second-order decision-analytic framework. The framework provides means for retrospective assessment of interim decisions based on a clear and consistent set of criteria that combines both ethical and epistemic considerations. The framework is broadly Bayesian and addresses a fundamental question behind many concerns about clinical trials: What does it take for an interim decision (e.g., whether to stop the trial or continue) to be a good decision? Simulations illustrating the modeling of interim decisions counterfactually are provided.* ***Key words:*** *ethical framework; interim analyses; stopping rules; statistical decisions; group sequential methods; DSMB; Bayesian; counterfactual reasoning.* ***(Med Decis Making 2016;36: 999–1010)***

What does it take for an interim decision (e.g., whether to stop the trial or continue) to be a good decision? This question is important not only to data and safety monitoring board (DSMB) members and ethicists but also to anyone who wants to be in a position to better understand interim decisions in clinical trials. Most approaches to monitoring and interim analyses do not discriminate effectively between good and bad decisions. While statistical approaches tend to focus on epistemic aspects of monitoring rules (cf. Goldman and Hannan,[1] Proschan and others,[2] and Gillen and Emerson[3]),[*] often overlooking ethical considerations, ethical approaches to randomized controlled trials (RCTs) tend to overlook key epistemic dimension (cf. Freedman[5]—the principle of clinical equipoise; Buchanan and Miller[6]—the principle of

[*] Proschan and others[2] is a good example of this disregard. Their unified approach aims at modeling the mechanism of accumulating data in clinical trials as following a single motion, which they call ''Brownian motion,'' where the test statistic and treatment effect estimator are treated as if they were a sum and mean, respectively, of independent and identically distributed random variables. Throughout their approach, they ''stress the need for statistical rigor in creating an upper boundary.'' While the ''upper boundary'' of monitoring rules aims at demonstrating the efficacy of new treatments, ''lower boundaries'' deal with harm, or ''unacceptable risk.'' Having said that, ''most of [their approach] deals with the upper boundary [problem] since it reflects the statistical goals of the study and allows formal statistical inference. But the reader needs to recognize that considerations for building the lower boundary (or for monitoring safety in a study without a boundary) differ importantly from the approaches to the upper boundary'' (p. 6). We find a similar approach also in Moyé.[4] Neither study devotes serious attention to safety, despite acknowledging its significance.

Received 30 April 2015 from Department of Philosophy, Ottawa Hospital Research Institute, University of Ottawa, Canada. Revision accepted for publication 9 May 2016.

Address correspondence to Roger Stanev, PhD, Centre for Practice-Changing Research, Ottawa Hospital Research Institute, and Department of Philosophy, University of Ottawa, 501 Smyth Rd, L1218, Ottawa, ON Ontario K1H 6K8, Canada; e-mail: rstanev@uottawa.ca.

DOI: 10.1177/0272989X16655346

nonexploitation).[†] The proposal is a quantitative framework—a second-order decision-analytic framework—that incorporates both ethical and necessary epistemic considerations for the modeling and retrospective assessment of interim decisions.

The 2 most basic tasks of the framework are to deliver the following:

1. A reconstructed justification for a given interim decision
2. A verdict as to whether the decision is ethically permissible

The article proceeds as follows. The first section sets out basic requirements and indicates some important restrictions on the framework. The second section gives an overview of the framework. The third section presents the formal elements constituting the framework and a simulation illustrating the modeling of interim decisions. The article concludes raising further questions of epistemological interest regarding the modeling and evaluation of interim decisions and the public accountability of such decisions.

## BASIC REQUIREMENTS AND RESTRICTIONS

Based on past DSMB experiences from the Women's Health Initiatives trials,[7–9] as well as human immunodeficiency virus (HIV)/AIDS and oncology trials,[10,11] the proposed framework should meet certain requirements. The most basic requirement (which underlies all others below) is the idea that every DSMB decision should articulate a clear general principle (i.e., a *ratio decidendi*) that gives reason for the interim decision, and all such decisions should be made public. This section sets out the basic requirements and indicates important restrictions.

### Clarity

The framework must provide clear standards of representation (modeling) and clear criteria of evaluation. We need to be clear both about how an interim decision is represented (modeled) and about the principles used for its evaluation.

### Scope

The guidelines for representation must be flexible enough to accommodate the wide variety of interim decisions in clinical trials. That is because different types of decisions reflect different types of dilemma, which consequently appeal to contextual factors differently. The framework must be broad enough to accommodate 1) the different types of early stopping decisions typically found in RCTs, 2) common disagreements about decisions based on them, and 3) common types of statistical monitoring procedures—whether based on frequentist or Bayesian philosophies.

### Flexibility

The framework should yield answers that vary with changes in problem representation. The framework assists with the identification of a set of relevant decision criteria and with representation of the monitoring rules that govern stopping decisions. Having identified these criteria, the framework tries to represent decisions as optimal, in the following weak sense of optimality: no alternative representation does better on every criterion.

### Explanatory Power

The framework should provide public criteria amenable to justification. An adequate framework should show how good interim decisions actually contribute to the permissibility of these decisions. Because a proper evaluation of a DSMB decision requires a reason for its decision, a good decision requires a fully structured justification capable of being communicated to others (i.e., a public rationale or *ratio decidendi*).[‡] The conception of a good decision proposed by the framework does not judge an interim decision as right or wrong based on whether the DSMB got the outcome (e.g., superior therapy) correctly or incorrectly. The criteria for a good decision focus instead on whether there is a principled reason for the interim decision and then whether the rational decision is ethical, under some ethical principle or maxim (e.g., minimax, since it

[†] With assessments grounded on the proportionality of the trial's estimated risks and potential benefits.

[‡] *Ratio decidendi* is a term that originates from the law. It means either "reason for the decision" or "reason for deciding." It is the public portion of the court's (judge-made law) ruling.

safeguarded trial participants by minimizing maximal losses).[§]

### Balancing Epistemic and Ethical Factors

Frameworks currently available tend to focus on the epistemic performance of monitoring rules—such as those in statistical literature—or on meeting the demands of medical ethics applied to clinical research, such as those proposed by the notion of "equipoise disturbed" (cf. Freedman[5]), "therapeutic obligation" (cf. Marquis[12]), "nonexploitation" (cf. Buchanan and Miller[6]), or "risk-benefit relationship as a moral compass" (cf. Iltis).[13] Yet such approaches lead to the view that interim decisions should be assessed on the basis of either "scientific" or "medical" objectives. The proposed framework, by contrast, represents decisions by combining objectives on the basis of factors contextualized in a particular decision situation.

To meet these requirements, the first step is to appeal to a basic classification of interim monitoring decisions. These are whether or not to stop due to efficacy, futility, or harm.[11] This classification permits the identification of relevant criteria that might influence the decision. The next step, given the identification of a set of criteria for representation, is to develop a specialized decision-analytic model, given the existence of a philosophical theory of rational decision making. This means defining the components of the decision-analytic model:

1. A set of available actions (at any interim point in the trial, e.g., stop or continue)
2. Foreseeable consequences of each action
3. Set of possible states of the world (i.e., a set of hypotheses about the effect of interest, e.g., null, alternative)
4. Set of alternative statistical monitoring rules
5. Decision criteria

*We call this a complete DSMB decision specification.*

A fundamental constraint on the decision specification is the *weak optimality* requirement: having identified a set of relevant decision criteria, concentrate on representations under which actual DSMB decisions are *weakly optimal*, in the sense that no alternative under consideration fares better on every criterion. Conflicts of expert opinion can then be analyzed by exploring the distinct ranges of assumptions (scenarios) under which conflicting stopping strategies emerge as optimal choices.

This broadly comparative Bayesian approach offers the hope of meeting all 5 requirements listed above.[**] Because we have the components of decision theory *within* the model, the framework allows us to explicitly account for the epistemic and ethical components that should influence the interim monitoring decision of clinical trials. By being able to represent interim decision options as decisions under risk, each interim decision will, in turn, depend on epistemic elements (e.g., estimated effect size, population frequencies or patient horizon, predictive probability of the test statistic given the stopping rule to reach certain stopping boundaries) and ethical elements (e.g., minimization of maximal losses, expected costs and benefits of possible treatment outcomes). In this light, the decision model can meet the balancing requirement of epistemic and ethical factors.

To conclude this preliminary section, we note 2 important restrictions to the decision-analytic model. The first is that the decision-analytic model is capable of representing only some of the criteria that might, in real life, influence data monitoring decisions. The second is that the framework does not aim to provide an algorithm for making early stopping decisions. It does not propose a theory that takes a description of an RCT decision as input and delivers a sharp verdict about whether or not the trial should be stopped. Instead, it offers a broad general framework that allows modelers to explore and analyze the justification for stopping decisions.

Moreover, the decision framework is not intended as a strict reconstruction of what DSMBs actually have done, or often do, when faced with the complexity of interim decisions. Deliberations from DSMBs are publicly unavailable. When the

[§] According to the framework, having a true favorable outcome in retrospect is neither a necessary nor sufficient condition for a good decision. Putting it bluntly, a good decision is having a fully structured justification for the decision and being justifiable (i.e., whether there is a principled reason for the decision), whereas a right (or wrong) decision is a matter of whether the decision simply followed from its monitoring rule. Every good DSMB decision articulates a clear and general principle (a *ratio decidendi*) that gives reason for the decision in the trial.

[**] Spiegelhalter and others[14] is another example of a Bayesian approach that attempts to combine epistemic and ethical considerations during the design, monitoring, and interim analyses of clinical trials.

deliberations do exist, they are often relegated to "Appendix 16.1.9" of reports submitted to regulatory authorities such as the US Food and Drug Administration (FDA).[††] In rare circumstances, there is an actual case study identifying and discussing "lessons learned" specific to the clinical trial (cf. DeMets and others[10]), and even then, relevant factors of interim decisions are often missing, making the public transcript an unreliable basis for anyone trying to appraise the DSMB's deliberations.

Given that the DSMB has an information monopoly during all interim analysis, also having sweeping discretion over the course of the trial precludes most meaningful oversight of its decision making.[15] Decision-making discretion by the DSMB becomes particularly challenging given the added fact that most of its deliberations happen behind closed doors, routinely not reporting publicly its interim decision reasons and recommendations.[16]

Although there are practical reasons for DSMBs to keep interim data analysis private under the premise of confidentiality, secret DSMB decision making has at least one important shortcoming: the lack of publicity in decision making prevents the public from getting a proper understanding of the reasons for the DSMB findings and final recommendation. Without a public rationale for its decisions (e.g., early stop, continue, changes to the trial), DSMB decision making prevents others from reaching their own conclusions about the trial's ethical and scientific appropriateness. And this is an important distinction from the way decision making by court of law happens, for instance, particularly in higher judicial decisions (e.g., setting a precedent) when a judgment is made explicitly and publicly with the inclusion of the judge's reasoning over the appropriate resolution of the legal issue.

Stanev[17] argues that decision making in legal systems such as judge-made law strikes an optimal balance between the competing demands of conservatism (with *stare decisis*, the rule that like cases should be decided alike) and innovation (the continuous development of the legal system). Based on similar relationships in the ways DSMBs rely on rules to make decisions in clinical trials, my argument by analogy had focused on conveying plausibility upon the need for publicity and explicitness in DSMB decision making—contrary to current, secretive DSMB practice. If the analogy succeeds, it shows that a similar explanatory hypothesis in clinical trials would explain a similar consequence: DSMB decision making striving for a balance between conservatism and innovation—avoiding dangerous medical treatments and bringing new and effective treatments into use as rapidly as possible—should promote the publicity and explicitness of decisions, the sort of public justification that the framework here proposes for decision making.

The framework begins with an interim decision stripped of some of its complex features. The simplification allows the decision analyst (i.e., modeler) to focus on a limited set of questions about the DSMB interim decision at stake. The particular factors and details that are not stripped vary with the specific type of decision (e.g., early stop due to efficacy, harm, futility). An alternative way of putting this point is that different factors are held fixed depending on the interim decision. This means that the first step of the framework is to provide a simplified description of an interim decision.

## OVERVIEW OF THE FRAMEWORK

The reconstruction of interim decisions requires certain elements. The reconstruction may include the actual stopping rule—if publicly available—or some reasonable reconstruction of the actual prespecified protocol stopping rule. The model is also meant to capture the reasoning that led to the selection of that (actual or reconstructed) stopping rule. This requires the representation of a set of alternative decisions (e.g., continuing with the trial at first interim or stopping), a set of expected losses, and one or more alternative statistical monitoring plans. The statistical monitoring plan includes the trial stopping rule, the number of interim analyses, and the efficacy on primary outcome in which the study was based, as reported in the study's original publication (or as specified in its protocol).

The representation is considered incomplete until the interim decision (e.g., early stopping) is justified as permissible. The standard for permissibility is the existence of at least one decision

[††] According to ICH E3, "Any operating instructions or procedures used for interim analyses should be described. The minutes of meetings of any data monitoring group and any data reports reviewed at those meetings, particularly a meeting that led to a change in the protocol or early termination of the study, may be helpful and should be provided in Appendix 16.1.9. Data monitoring without code-breaking should also be described, even if this kind of monitoring is considered to cause no increase in type I error" (p. 21).

criterion on which the interim decision is "optimal" (i.e., it maximizes expected utility). We hinted at the meaning of optimality earlier; a little more about this idea is in order.

According to the framework, a decision criterion prescribes a way in which, for a given representation of an early stopping decision, the criterion picks an "optimal stopping decision." Because criteria are given by the type of decision (or dilemma) and salient factors of the case, the decision criteria are relative to a set of predefined criteria. We say that a decision is permissible if it is "optimal" according to at least one decision criterion. If we succeed in giving the RCT case a representation and show that the stopping decision is "optimal" on at least one decision criterion, we then say that the decision is permissible, or in principle justifiable.

The last step of the framework is evaluation. Evaluation is composed of 2 stages: a policy stage and a running stage. The first stage reflects a policy decision that is logically prior to the running stage (any data collection or conduct) of the RCT. The first stage is concerned with the choice of the monitoring rule per se (i.e., with the justification of a choice of monitoring rule). The second stage is concerned with the justification of a particular action (i.e., decision) falling under its stopping rule. This second stage may include the "optimal" decision criterion if one is identified during the reconstruction of the early stopping decision.

The need for a 2-stage evaluation is supported by appeal to an intuitive understanding of what ethics (i.e., a just or fair assessment of decisions) requires. An example serves to illustrate the intuition. We seem to give more weight to harm when it is engendered by a rule that is officially adopted by a DSMB than when the harm is due to insufficiently enforced rules or insufficiently protective DSMBs. For instance, burdens on trial participants due to an authorized restriction (e.g., a highly stringent and unrealistic stopping condition) seem morally more serious than burdens engendered by poorly monitored RCTs (e.g., ignoring unforeseen effects) or by DSMB skeptics (e.g., skepticisms about the clinical importance of certain interim results), even if the harms on trial participants are *exactly* the same—or if the probability of harm faced by trial participants is the same. Unless our criteria of judging the permissibility of a case save this intuitive distinction, an ethical evaluation of interim decisions is incomplete. Thus, we need to make a distinction between the choice of monitoring rule and the decision about whether to stop or continue given a particular stopping rule.

To accommodate this distinction between the choice of a monitoring rule and the choice of a particular action falling under it, our evaluations of monitoring decisions must not simply compute or "tally up" in some way the foreseeable effects each available action would have on trial participants. When judging permissibility, a more complete evaluation of an early stopping decision must weigh these effects differently for distinct types of a causal link connecting the candidate decision to benefits and harms for trial participants. And this can be accomplished with the 2-stage process outlined above.

Consider the following claim: "The DSMB stopped its trial early due to efficacy." Until the reader knows the type of trial concerned and its governing rules—rules that defined the acts and conditions of early stopping—she or he does not fully understand what the DSMB did. The 2-stage evaluation is necessary not only to account for the fact that exceptions to stopping rules are not uncommon in practice but also because a breach of the stopping rule by the DSMB, which may appear to maximize expected utility according to some representation, may not in fact have an adequate justification. The fact that a decision to continue a trial—increasing the chances of harm to trial participants—appears to be justified as maximizing expected utility *given the choice of stopping rule* should not be considered an adequate justification—unless the rule under which the action falls is also appropriate to the given RCT context.[‡‡]

Another point about the need for a 2-stage evaluation is in order. Even though there is a commitment always to represent the DSMB decision as "optimal" (permissible), the reconstruction of an early stopping decision, by itself, does not imply having a complete evaluation of the early stopping decision. The reason is this: in cases where there are rival opinions about the appropriate stopping rule, it may be demonstrated *by comparison* that a rival stopping rule is better suited or more appropriate given the RCT context.

By using the framework, decision analysts are able to compare alternative DSMB decisions. Why

[‡‡] This distinction is motivated by "two concepts of rules" (see Rawls[18]). The distinction is further motivated by an "intuitive understanding of justice," as seen in Pogge[19] when critically examining Rawls.[20]

did the DSMB choose to continue rather than to stop the trial? Given the adoption of a particular statistical monitoring rule, had the foreseeable consequences of its actions been different, would the DSMB have taken a different course of action? Under what realization of expected losses would the DSMB have chosen a different course of action? These counterfactual ''what-if'' questions reveal a conscious attempt on the part of the reader to provide reasons for the DSMB decision. This point about counterfactual questions indicates a counterfactual-style intuition about explanations—namely, that we have explanations if we can answer counterfactual questions.[§§] The suggestion is that what-if-things-had-been-different information is intimately connected to our judgments of explanatory relevance.

A distinguishing feature of this counterfactual style of explanations is that they are explanations that furnish information that is potentially relevant to control and manipulation. These counterfactual explanations inform us, decision analysts, when we are able to change the value of one or more variables (e.g., monitoring rule R1 to R2, ethical loss function to scientific loss function, or changes in utility values), we could change the value of other variables (e.g., continue the trial v. stop the trial). This conception of explanation has the advantage of fitting a wide range of medical science contexts, particularly clinical trial monitoring and behavioral sciences, where investigators think of themselves as revealing patterns of regularities and constructing explanations.

A key virtue of such explanations is that they show how what is explained depends on other, distinct factors, where the dependence in question has to do with some relationships that hold as a matter of manipulability via a model, which is very useful for decision analysts and those modeling reasoning. This way, decision analysts, by manipulating their model, can see if the model can account for potential, yet plausible, differences of opinions among DSMB members. If it can, we have provided a plausible explanation for reasonable disagreements among expert opinions. If it cannot, then we still have a problem, which is to account for expert opinions that can (might) differ for good reasons.

The idea worth pointing out here is that the decision analyst gains understanding of the DSMB's interim decision in the ability to grasp a pattern of counterfactual dependence associated with relationships that are potentially exploitable for purposes of manipulation via a model. For instance, we might come to understand how changes in patient health status will change the conservatism of early stopping rules: as health increases, greater efficacy is needed to outweigh possible long-term side effects. Grasping such relationships is the essence of what is required for the analyst to be able to explain the DSMB's decision to continue the trial despite early evidence for efficacy.[***]

## FORMAL ELEMENTS AND SIMULATION

According to the framework, the representation of a statistical monitoring plan begins as a decision under uncertainty, represented as a 4-tuple ($\theta$, $A$, $Y$, $L$). $\theta$ is the parameter space (the set of possible true states of nature, $A$ is the set of possible actions, $Y$ is the observation (data) model, and $L$ is a specific loss function. Therefore, $\delta(y)$: $Y \rightarrow A$ is a statistical monitoring rule.

To illustrate the framework for modeling decisions, the example is kept as simple as possible. Approximate analysis that considers relevant features shall suffice for the purposes of this article. We treat the progression to AIDS as a binary event. The example involves dichotomous observations, single pair of treatments, 2 hypotheses, 1 loss function, 2 statistical monitoring rules, 1 ''overarching'' maxim, 3 types of actions, and easy to compute numbers.

Suppose $N$ individuals have HIV. There are 2 treatments: standard $T_1$ and experimental $T_2$. To

[§§] ''We see whether and how some factor or event is causally or explanatorily relevant to another when we see whether (and if so, how) changes in the former are associated with changes in the latter.''[21]

[***] This is an important view of explanation and agency that goes all the way back to German philosopher Immanuel Kant. It distinguishes causal descriptions from actions. Consider the contrast between actions and other occurrences that do not involve self-directed thought and agency. Kant said that all events happen for reasons, but only actions are done for reasons. Only actions can be explained by reference to the agent's (e.g., DSMB) reasons, that is, considerations the DSMB itself took to be normative reasons weighing in favor of its choice. In this manner, the framework follows an ''ethics of expertise'' principle that has been articulated by philosophers of science—namely, when value judgments influence scientific decisions, scientists should make those influences as explicit as possible[22] as part of their justifications.

find out which of the 2 treatments is more effective, an RCT is conducted on $2n$ of the total $N$ patients, with $n$ patients assigned to each treatment. If the trial ends, the remaining $N - 2n$ patients receive the treatment selected as the more effective, unless no treatment is declared superior, in which case the remaining patients are treated with standard treatment $T_1$.

We restrict to statistical monitoring rules that permit termination after $n/2$ (half the number of participants assigned to each treatment group) have been treated. That is, the RCT has a single interim analysis halfway through the trial. Keeping calculations relatively simple, $N = 100$, and of these, we assume the monitoring of 10 + 10 patients ($n = 10$) so that 5 patients assigned to each treatment are treated by the interim analysis, and 10 patients assigned to each treatment are treated by final analysis.

### Observation Model

For each treatment, a single outcome is observed: whether or not the patient recovers. The probability of recovery with $T_i$ is $\theta_i$ ($i = 1, 2$) (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. The observations are modeled as to follow a binomial distribution:

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

Since the difference between $T_1$ and $T_2$ is of interest, let $\theta$ stand for the difference in response rates between $T_1$ and $T_2$ so that $\theta = \theta_2 - \theta_1$. Thus, the joint distribution of $y_1$ and $y_2$ given $\theta_1$ and $\theta_2$ is

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

where $y$ and $\theta$ denote $(y_1, y_2)$ and $(\theta_1, \theta_2)$, respectively; $y \in Y$ and $\theta \in \theta$.

### Parameter Space θ

The model assumes that the recovery rate of $T_1$ ($\theta_1$) is 0.5 and that of $T_2$ ($\theta_2$) is unknown. It assumes that $\theta$ takes 1 of 2 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., $\theta_2 = 0.7$).

### Set *A* of Allowable Actions

At the interim analysis, the DSMB can act in 1 of 3 ways:

$a_1$: Stop and declare that "$T_2$ is more effective than $T_1$,"
$a_2$: Stop and declare that "$T_2$ is equal to $T_1$," or
"recommend a continuation of the trial."

If DSMB continues the trial at interim, then at the end of the trial, it can act in 1 of 2 ways:

$a_{1.f}$: Stop and declare that "$T_2$ is more effective than $T_1$,"
$a_{2.f}$: Stop and declare that "$T_2$ is equal to $T_1$."

At either the interim 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 $(y_1, y_2)$ lead to the choice of action to be taken—a choice that depends ultimately on the decision monitoring rule. A decision monitoring rule is a function of sample data $y$ leading into the set $A$ of allowed actions $\{a_1, a_2, a_{1.f}, a_{2.f}\}$. It specifies how actions are chosen, given observation(s) $y$.

### Decision Stopping Rule $\delta(y)$: $Y \rightarrow A$

We first consider decision monitoring rule $R_1$. This rule is based on frequentist properties (Neyman-Pearson inference), where the focus is on controlling error probabilities. We use an instance of group sequential monitoring procedure, much like an O'Brien-Fleming stopping rule.[23] In this stopping rule, $R_1$ aims at controlling the overall type I error by having it be no more than 0.05 (approximately). For this, the DSMB considers rejecting $H_0$ (if $H_0$ is true) for values $(y_2 - y_1) \geq 4$[†††]; otherwise, the DSMB does not reject $H_0$.

At interim (when $n = 5$):

$a_1$: Reject $H_0$ (assume $H_0$ is false) for values $(y_2 - y_1) \geq 4$, i.e., for $(y_2 - y_1) = \{4, 5\}$; stop and declare "$T_2$ is more effective than $T_1$";
$a_2$: Accept $H_0$ (assume $H_0$ is true) for values $(y_2 - y_1) \leq -4$, i.e., for $(y_2 - y_1) = \{-4, -5\}$; stop and declare "$T_2$ is equal to $T_1$."

[†††] $P((y_2 - y_1) \geq 4 | H_0) = 0.05$, computed according to the joint probability distribution. For example, $P((y_2 = 7, y_1 = 3) | H_0) = (0.117)^2 = 0.014$.

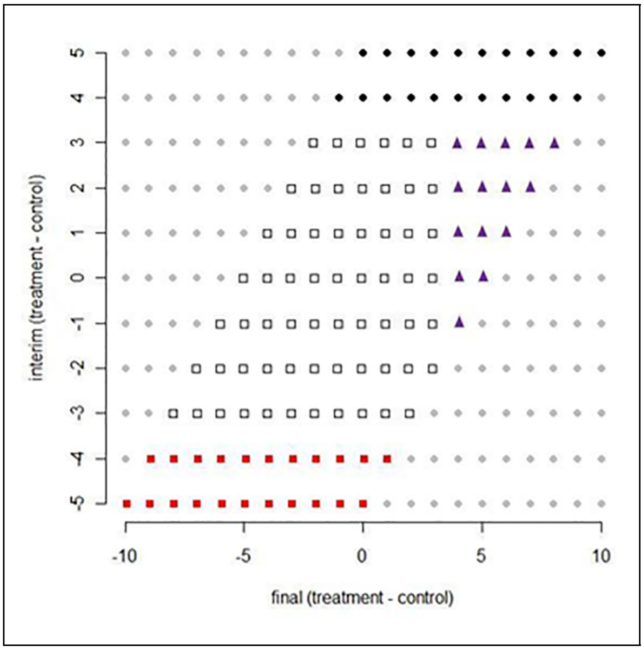

*Figure 1 Decision rule $R_1$ as a function of possible values of* ($y_2$ – $y_1$)*, with* ($y_2$ – $y_1$) *at the end of the trial plotted as a function of* ($y_2$ – $y_1$) *at interim. Actions: $a_1$ (black circles), $a_2$ (red squares), $a_{1\cdot f}$ (purple triangles), and $a_{2\cdot f}$ (white squares).*

This means that for $(y_2 - y_1) = \{-3, -2, -1, 0, 1, 2, 3\}$, the DSMB continues with the trial.

At the end of the trial (when $n = 10$):

$a_{1.f}$: Reject $H_0$ (assume $H_0$ is true) for values $(y_2 - y_1) \geq 4$, i.e., for $(y_2 - y_1) = \{4, 5, 6, 7, 8\}$; stop and declare "$T_2$ is more effective than $T_1$"; otherwise,

$a_{2.f}$: Do not reject $H_0$ for $(y_2 - y_1) = \{-8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3\}$; declare "$T_2$ is equal to $T_1$."

Figure 1 shows how $R_1$ works as a function of possible values $(y_2 - y_1)$, both at interim and at the end of the trial. Specifically, the figure plots the range of $(y_2 - y_1)$ values *at the end of the trial*—indicated by the x-axis, that is, *final* (treatment − control)—conditional on values $(y_2 - y_1)$ *at interim*—indicated by the y-axis, that is, *interim* (treatment − control). $a_1$ is represented by the set of black circles, $a_2$ is represented by the set of red squares, $a_{1\cdot f}$ is represented by the set of purple triangles, and $a_{2\cdot f}$ is represented by the set of white squares. Gray circles represent impossible values for $(y_2 - y_1)$. For example, coordinate (X = 6, Y = 0) is not an option, given that $(y_2 - y_1) = 0$ at interim eliminates the possibility of $(y_2 - y_1) = 6$ at the end of the trial.

For the alternative monitoring rule (i.e., $R_2$), we assume a simple Bayesian monitoring rule for comparison. $R_2$ 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": $p(\theta = 0) = p(\theta = 0.2) = 0.5$. 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) to produce the ratio of posterior probabilities. Therefore, when discriminating between $H_0$ and $H_1$ on the basis of $y$, $R_2$ chooses the hypothesis with the larger posterior probability. For instance, putting it in terms of rejecting $H_0$, $R_2$ can reject $H_0$ when the likelihood ratio is less than 1. If so, with flat priors, $R_2$ rejects $H_0$ as long as $y_2 \geq 6$; otherwise, it does not reject $H_0$.

However, to make the 2 decision rules comparable despite their different statistical philosophies, we chose a cutoff point during the interim analysis—which is my point of contention—so that $R_2$ would be as close as possible to $R_1$ on this epistemic factor, while keeping with our choice of easy to compute numbers. Therefore, the sum of the probabilities, whose outcome's likelihood ratio is more extreme than the 3:1 or 1:3 cutoff, brings us to 0.04, which is the closest the rule gets to the type I error (0.05) used by the DSMB with $R_1$, given our choice of $n = 5$ at interim. Figure 2 shows how Bayesian monitoring rule $R_2$ works as a function of possible values $(y_2 - y_1)$, both at interim and at the end of the trial. One relevant difference between Figure 1 and Figure 2 is that, because the evidence provided by data $y$ is contained in the likelihood ratio—given that $R_2$ is a Bayesian monitoring rule—the probability values of the control group are irrelevant. They cancel out in the ratio because $\theta_1 = 0.5$, regardless of whether $H_0$ is true or $H_1$ is true. Only $\theta_2$ varies—namely, the treatment's effect.

$$\frac{p(H_0|y)}{p(H_1|y)} = \frac{p(H_0)p(y|H_0)}{p(H_1)p(y|H_1)}$$
$$= \frac{p(H_0)p(y_1|H_0)p(y_2|H_0)}{p(H_1)p(y_1|H_1)p(y_2|H_1)} = \frac{p(y_2|H_0)}{p(y_2|H_1)}$$

when priors $p(H_0) = p(H_1)$.

## Loss Function

This function is meant to capture ethical and epistemic values associated to possible outcomes of the trial. Following Heitjan and others[24] on loss functions, we start with the "ethical" loss function $L_E(\theta, \mathbf{a})$. $L_E(\theta, \mathbf{a})$ can be used to compare the 2 treatments by *paying a penalty for each patient assigned the inferior treatment.* One penalty point is assessed for each patient assigned the inferior treatment. One way to think of $L_E(\theta, \mathbf{a})$ is to approach it with respect to a particular patient. If $H_0$ is true, and if the patient is given $T_1$, the loss incurred with such treatment is 0 (since $T_1$ and $T_2$ are considered equivalent), but if the patient is given $T_2$ in this situation, then there is a cost, which we refer to as *cc*.[‡‡‡] The other possibility is that $H_1$ is true. In this case, if the patient is given the inferior treatment $T_1$, the loss is ***d*** (which we assume for now is 1 unit); if given $T_2$ (superior treatment), the loss is 0.[§§§] Table 1 summarizes this situation.

Now we make $L_E(\theta, \mathbf{a})$ sensitive to the "effect size" $|\theta_2 - \theta_1|$. By "effect size," we mean the percentage difference of no progressions 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 expected (by association) from 1 fewer patient having a positive no progression. That is 0.2 of 5 patients, which equals 1; it is the result that DSMB members should have obtained (or could have expected) had they continued with the trial. Table 2 presents sample losses according to this version of $L_E(\theta, \mathbf{a})$, accounting for effect size, $n$, and $N$, for values of $\boldsymbol{d} = 1$ and $cc = 0.01$.

## "Overarching" Maxim

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

$$E_Y[L(\theta,\delta(y))|\theta] = \sum_{y_i \epsilon Y} L(\theta,\delta(y)) f_Y(y_i|\theta).$$

If we have prior information about θ, which can be expressed in terms of a prior probability mass

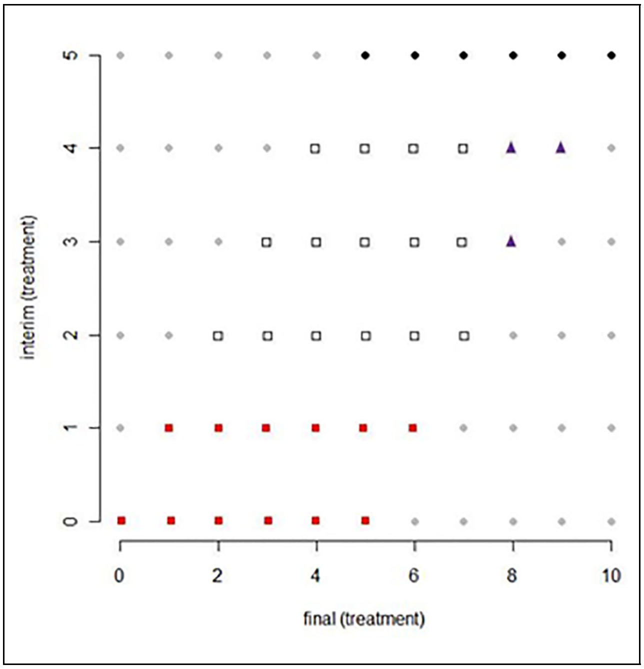

*Figure 2* *Decision rule $R_2$ as a function of possible values of $y_2$, with $y_2$ at the end of the trial plotted as a function of $y_2$ at interim. Actions: $a_1$ (black circles), $a_2$ (red squares), $a_{1\cdot f}$ (purple triangles), and $a_{2\cdot f}$ (white squares).*

**Table 1** Ethical Loss

| | $H_0$ | $H_1$ |
|---|---|---|
| $T_1$ | 0 | ***d*** |
| $T_2$ | *cc* | 0 |

**Table 2** Ethical Loss w/ Effect Size

| | **True State of Nature** | |
|---|---|---|
| **Action** | $\theta_2 - \theta_1 = 0$ | $\theta_2 - \theta_1 = 0.2$ |
| At interim ($n = 5$) | | |
| $a_1$: choose $T_2$ | $(N - n)$<br>$cc = 0.95$ | $(\theta_2 - \theta_1)$<br>$(n\boldsymbol{d} + (N - 2n)0) = 1$ |
| $a_2$: choose $T_1$ | 0 | $(\theta_2 - \theta_1)$<br>$(n\boldsymbol{d} + (N - 2n)\boldsymbol{d}) = 19$ |
| At final ($n = 10$) | | |
| $a_{1.f}$: choose $T_2$ | $(N - n)$<br>$cc = 0.9$ | $(\theta_2 - \theta_1)$<br>$(n\boldsymbol{d} + (N - 2n)0) = 2$ |
| $a_{2.f}$: choose $T_1$ | 0 | $(\theta_2 - \theta_1)$<br>$(n\boldsymbol{d} + (N - 2n)\boldsymbol{d}) = 18$ |

[‡‡‡] *cc* is a constant loss that can be conceptualized in different ways. For my purposes, it can be understood as the "cost" of having a patient subjected to a new treatment, when that treatment is no better than standard treatment.

[§§§] The "losses" can also be understood as "regrets" since they are the loss over and above the losses incurred under complete information (i.e., when every patient is managed the way that is optimal for her or him). I thank one of the reviewers for this point.

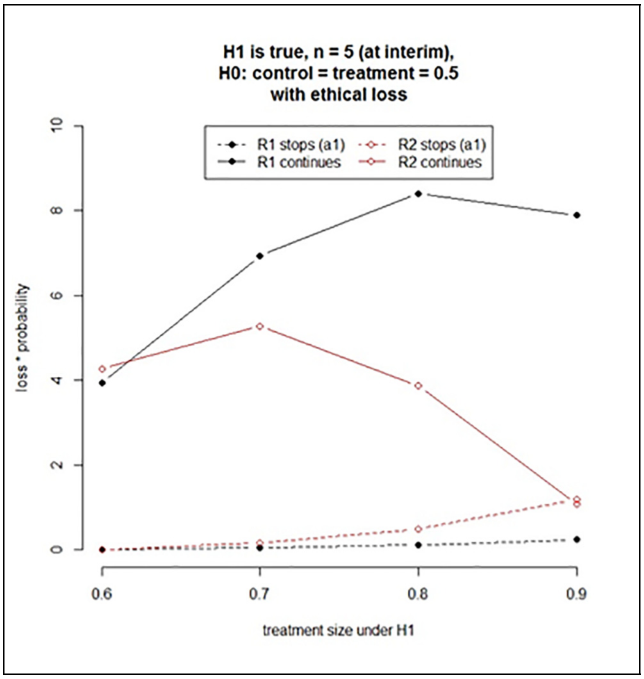

*Figure 3 Weighed losses of stopping (action $a_1$) and continuing (averaging actions $a_{1.f}$ and $a_{2.f}$) for stopping rules $R_1$ and $R_2$, when $H_1$ is true.*

function $p(\theta)$, then the Bayes risk of decision rule $\delta(y)$ is the expectation of the average loss over possible values of $\theta$:

$$r(\delta(y))=\sum_{\theta\epsilon\Theta}\sum_{y\epsilon Y}L(\theta,\delta(y))f_Y(y|\theta)p(\theta).$$

## Simulation

Figure 3 plots, while varying effect sizes, the weighted losses[****] for stopping and continuing, according to $R_1$ and $R_2$ decision procedures, when $H_1$ is true. If one assumes that the DSMB principle for stopping its trial is based on whether stopping it had a lower expected loss than continuing it (where the expectation is with respect to the weighted losses of continuing in light of the fixed set of future actions—$a_{1.f}$ or $a_{2.f}$), then, by averaging the weighted losses of $a_{1.f}$ and $a_{2.f}$, one can compare the weighted losses of stopping v. continuing for a whole range of effect sizes. Notice that for every effect size in the range, the DSMB decision to stop at interim, according to $R_1$, has a lower expected loss than continuing. The decision to stop is therefore consistent under $R_1$ given the range of effect size under consideration. This, however, is not the case with alternative decision rule $R_2$. Given the choice of ethical loss function and assuming that the principle to stop the trial following $R_2$ is also based on whether stopping the trial had a lower expected loss than continuing it, the only effect size that can warrant a decision to continue is a high expected treatment size (i.e., $\theta_2$ = 0.9). Otherwise, the trial should stop if following $R_2$ since stopping at interim had a lower expected loss than continuing.

If we focus our attention on the decision to continue the trial ($R_1$ continuing v. $R_2$ continuing), for the range of effect sizes, assuming an equal set of priors for every pair of hypotheses (i.e., each pair, $H_0$ v. $H_1$, having a different effect size for each $H_1$), $R_1$ is outperformed by $R_2$ with respect to the weighted losses except when the effect size is small, $\theta_2$ = 0.6. This is shown in the upper part of Figure 3. With respect to the weighted losses, $R_1$ is therefore "less" ethical (has greater weighted losses) than $R_2$ when the effect sizes are not small, that is, for values $\theta_2$ = {0.7, 0.8, 0.9}. This difference, however, may be of no surprise given the fact that $R_2$ is a Bayesian monitoring rule. A different set of priors would produce different weighted losses. This result is corroborated by a subsequent Bayes risk comparison.

Comparing $R_1$ against $R_2$ with respect to the Bayes risk, we see that the $R_1$ rule is "less" ethical (has a greater Bayes risk) than $R_2$'s when the effect sizes are not small, that is, for values $\theta_2$ = {0.7, 0.8, 0.9}. That is, for all effect sizes, except when $\theta_2$ = 0.6, the Bayes risk for $R_2$'s rule is smaller than the $R_1$ rule.

If, however, holding everything else fixed, we vary instead the loss function, we notice how the situation can be reversed. That is, if we shift from the "ethical" to a "scientific" loss function,[††††] the results are reversed. The comparison between the $R_1$ and $R_2$ rules is illustrated by the graphs in

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

[††††] The "scientific" loss function represents the idea that utility attaches only to finding the true state of nature, ignoring all other consequences. In particular, this loss function ignores the effects of mistreatment. From the point of view of the scientific loss function, correctly declaring that "$T_2$ is more effective than $T_1$" has the same utility as correctly declaring that "$T_2$ is equal to $T_1$," even though the 2 states of nature may have quite distinct consequences for present and future patients. In other words, the function assigns the same penalty for any incorrect conclusion. To be definite, we represent the penalty as a 10-unit loss. Table 3 summarizes this loss function.

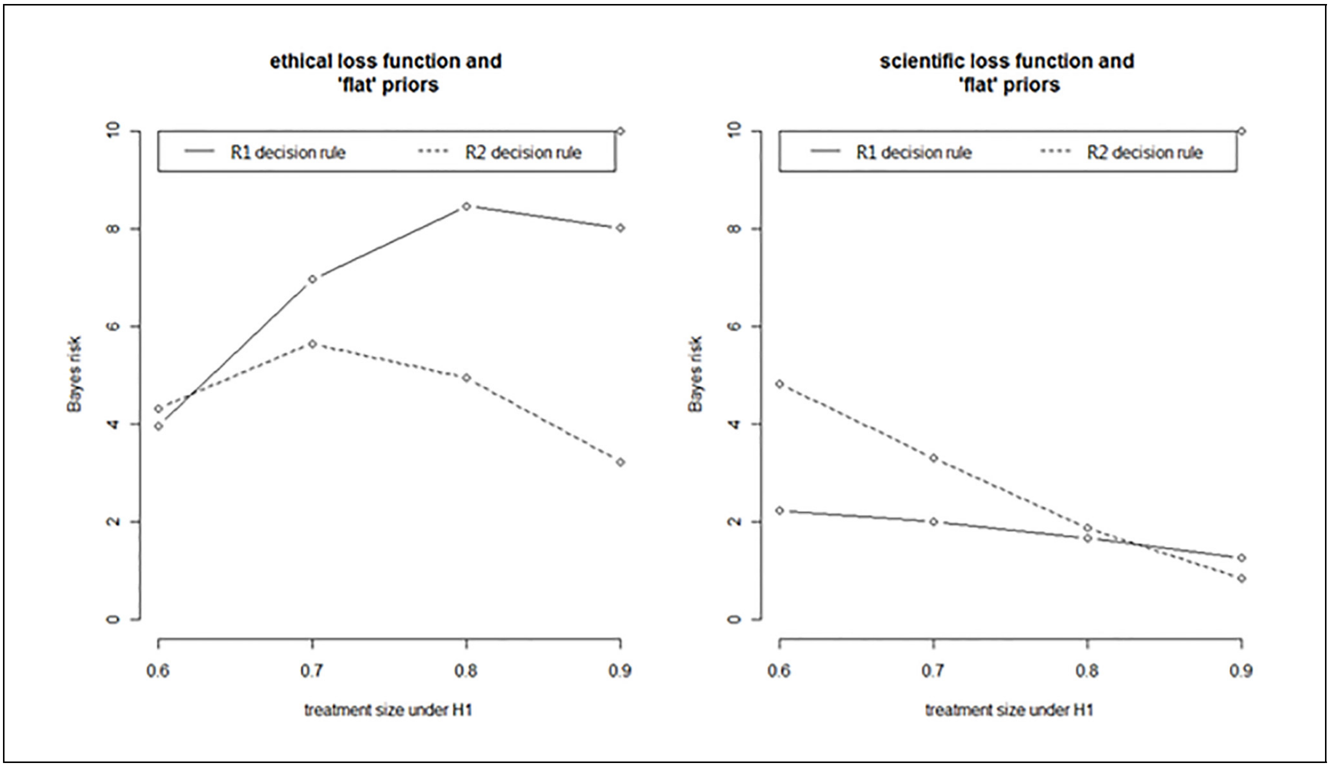

*Figure 4 Stopping rule $R_1$ v. stopping rule $R_2$ compared with respect to Bayes risk and different loss functions (ethical loss v. scientific loss).*

**Table 3** Scientific Loss

| | $H_0$ | $H_1$ |
|---|---|---|
| $T_1$ | 0 | 10 |
| $T_2$ | 10 | 0 |

Figure 4. In the graph on the right, the $R_1$ monitoring rule outperforms the $R_2$'s rule with respect to the Bayes risk in all effect sizes, except when effect is large, $\theta_2 = 0.9$. With respect to this function and using Bayes risk as a comparative rationale for adopting a monitoring policy, the $R_1$ rule is therefore overall "more" ethical then $R_2$'s except when the effect is large (i.e., $\theta_2 = 0.9$).

## CONCLUSION

The proposed framework for reconstructing interim decisions, although limited, systematizes reconstructions (in a plausible fashion), making implicit assumptions for interim decision explicit. The fact that the results of RCTs are often provided without public justifications for their interim decision is an important indictment of DSMBs' early stopping decisions.[23,25,26] If the proposed framework succeeds in providing the means to plausible reconstructions, then justifications for competing decisions exist. Without justifications, the generalizability of DSMBs' decisions and the subsequent analysis of their decisions are problematic. By providing a common framework that structures justifications in individual cases, one does indeed have the beginnings of a general model for interim decision making.

In pressing a systematic form of justification to which trialists either explicitly or implicitly subscribe, we, decision analysts, may have the means to regard whether a particular interim decision is either permissible or not, according to a given representation of the decision. Because it is not surprising that people often disagree on what is of greatest value or harm for them and for others, it is reasonable to seek an account of interim decisions that saves disagreements, without having to necessarily solve them once and for all. Moreover, when specifying a rationale for an early stopping decision, in a particular context, although different contexts may allow for comparisons between different interim decisions, judgments of superiority cannot be justified as always universally optimal—thus the need for a notion of weak optimality, as we have provided here, where one decision looks optimal in one representation—one set of dimensions—but not necessarily so on another.

Once the point about the balancing of epistemic and ethical factors has been appreciated, the following step is to consider what an early stopping principle of an RCT might look like. If principles of early stopping are to help guide DSMBs' behavior, these principles must be represented, and they must be capable of being communicated and taught so that they can serve as a public rationale for evaluating early stopping decisions. Any scientific (or ethical) approach whose principles of inference imply that it is not permissible to teach permissible interim decisions, or early stopping principles, violates this publicity standard—that is, it does not meet the public rationale for the early stopping of RCTs. In this fashion, it seems reasonable to say that in contrast to the 2 standard approaches to interim decision (either solely statistical or ethical), the proposed framework seems capable of meeting this publicity standard.

## ACKNOWLEDGMENTS

I am indebted to my colleagues at the Ottawa Hospital Research Institute, as well as Marc Levine, Paul Bartha, Hubert Wong, Alan Richardson, and John Beatty. Earlier versions of this work were presented at different conferences, including "New Science—New Risks" at the Center for Philosophy of Science, University of Pittsburgh, of which I am also grateful. I also thank the 3 reviewers for their helpful comments. Financial support through the Banting Fellow award from the Canadian Institutes of Health Research with the University of Ottawa and the Ottawa Hospital Research Institute is also gratefully acknowledged.

## REFERENCES

1. Goldman AI, Hannan P. Optimal continuous sequential boundaries for monitoring toxicity in clinical trials: a restricted search algorithm. Stat Med. 2001;20:1575–89.

2. Proschan MA, Lan KK, Wittes JT. Statistical Monitoring of Clinical Trials. New York: Springer; 2006.

3. Gillen DL, Emerson SS. Designing, monitoring, and analyzing group sequential clinical trials using the RCTdesign Package for R. In: Fleming TR, Weir BS, eds. Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials. New York: Springer; 2013. p 177–207.

4. Moyé L. Statistical Monitoring of Clinical Trials. New York: Springer; 2006.

5. Freedman B. Equipoise and the ethics of clinical research. N Engl J Med. 1987;317:141–5.

6. Buchanan D, Miller FG. Principles of early stopping of randomized trials for efficacy: a critique of equipoise and an alternative nonexploitation ethical framework. Kennedy Inst Ethics J. 2005; 15(2):161–78.

7. Anderson GL, Kooperberg C, Geller N, Rossouw JE, Pettinger M, Prentice RL. Monitoring and reporting of the Women's Health Initiative randomized hormone therapy trials. Clin Trials. 2007;4:207–17.

8. Wittes J, Barrett-Connor E, Braunwald E, et al. Monitoring the randomized trials of the Women's Health Initiative: the experience of the Data and Safety Monitoring Board. Clin Trials. 2007;4:218–34.

9. Goodman S. Rashomon revisited: two view of monitoring the Women's Health Initiatives trials. Clin Trials. 2007;4:205–6.

10. DeMets DL, Furberg C, Friedman L. Data Monitoring in Clinical Trials. New York: Springer; 2006.

11. Ellenberg SS, Fleming TR, DeMets DL. Data Monitoring Committees in Clinical Trials. New York: John Wiley; 2003.

12. Marquis D. Leaving therapy to chance. Hastings Center Rep. 1983;13:40–47.

13. Iltis A. Stopping trials early for commercial reasons: the risk-benefit relationship as a moral compass. J Med Ethics. 2005;31: 410–4.

14. Spiegelhalter D, Abrams K, Myles J. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. New York: John Wiley; 2004.

15. Eckstein L. Building a more connected DSMB: better integrating ethics review and safety monitoring. Accountability Res. 2015;22:81–105.

16. Wittes J. Behind closed doors: the data monitoring board in randomized clinical trials. Stat Med. 1993;12:419–24.

17. Stanev R. Data and Safety Monitoring Board and the ratio decidendi of the trial. J Philosophy Sci Law. 2015;15:1–26.

18. Rawls J. Two concepts of rules. Philosophical Rev. 1955; 64(1):3–32.

19. Pogge T. John Rawls. Oxford, UK: Oxford University Press; 2007.

20. Rawls J. Some reasons for the maximin criterion. Am Econ Rev. 1974;64:141–6.

21. Woodward J. Making Things Happen. Oxford, UK: Oxford University Press; 2003.

22. Douglas H. Science, Policy, and Value-Free Ideal. Pittsburgh, PA: University of Pittsburgh Press; 2009.

23. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;35:549–56.

24. Heitjan D, Houts P, Harvey H. A decision-theoretic evaluation of early stopping rules. Stat Med. 1992;11:673–83.

25. Moher D. The CONSORT statement. BMC Med Res Method. 2001;1:2.

26. Mills E, Cooper C, Wu P, Rachlis B, Singh S, Guyatt GH. Randomized trials stopped early for harm in HIV/AIDS: a systematic survey. HIV Clin Trials. 2006;7(1):24–33.