JOHNS HOPKINS
UNIVERSITY

Center for Clinical Trials

Department of Biostatistics
Department of Epidemiology
Department of International Health

Department of Medicine
Department of Ophthalmology
Oncology Center

(Sunday) 4 March 2001

Memorandum

To: Center for Clinical Trials faculty and staff

Fr: Curt Meinert

Re: Reasons to be wary stopping rules

Definitions

data dredging *v* - **Data analyses** done on an **ad hoc** basis, without benefit of prior stated **hypotheses**, especially those done with the aim or intent of trying to find noteworthy differences within or among different **subgroups**; **exploratory data analysis**; see **dredge**. ***Usage note***: Often used in a pejorative sense, especially in reference to analyses in which it appears that only large differences are presented and where the number of **comparisons** made is not specified. Not to be confused with **subgroup analysis**; see usage note for that term.

design variable *n* - [**trials**] The **variable** used for determining or justifying **sample size** in planning a trial. rt: **primary outcome variable** ***Usage note***: Not to be used interchangeably with **primary outcome**, **endpoint**, or **primary endpoint**. Generally, the design variable denotes an important measure in the context of a trial, but it is rarely the sole measure of interest. See **primary outcome** for additional comments.

early stop *n* - [**trials**] An instance in which a trial is stopped prior to its scheduled end, especially because of accumulated **data** from within the trial suggesting **benefit** or harm associated with one of the **study treatments**. syn: **premature stop** ***Usage note***: See comment for **early stopping** and **trial stop**.

efficacy monitoring *v* - [**trials**] 1. **Monitoring** (defn 2) for **efficacy**, as performed at periodic time points over the course of a trial, to determine whether the trial should be stopped or modified; as distinct from **safety monitoring**. 2. **efficacy review** (defn 1) 3. **treatment effects monitoring** rt: **safety monitoring**, **treatment effects monitoring**, **interim look**, **interim result** ***Usage note***: Often used in contradistinction to **safety monitoring** in settings where the user wishes to distinguish between **interim looks** performed for efficacy monitoring versus those made for safety monitoring; eg, in settings where looks for safety monitoring are not counted as looks for purposes of adjusting **p-values** for **multiple looks**. The distinction is predicated on the assumption that safety and efficacy are **independent** dimensions of **treatment** — often not the case. Use **treatment effects monitoring** when the distinction is unimportant or where the monitoring performed is for efficacy and safety. See also notes for **administrative review**, **safety monitoring**, and **treatment effects monitoring**.

group sequential *adj* - 1. Of, relating to, or concerned with a **sequential** process in which the **unit** defining the **sequence** is an **aggregate** of **observations** or **treatment units**. ant: **unit sequential** 2. Relating to or based on a method of **hypothesis testing** involving use of accumulating data, augmented before each new **test** by an added set of **observation** (as in **group sequential interim data analysis**); the testing process continues until the **null hypothesis** is rejected, or some other **boundary** condition is encountered or crossed.

group sequential design *n* - [**trials**] A **sequential design** (**open** or **closed**) in which **treatment comparisons** are made only at designated time points or intervals (eg, every 6 months), after **enrollment** of specified numbers of **patients** (eg, after every 50 **randomizations**), or after occurrence of specified numbers of **events** (eg, after every 10 deaths) and the results of which are used to decide whether or not to continue the trial. ***Usage note***: See **sequential data analysis**.

group sequential interim data analysis *n* - [**trials**] A method of **interim data analysis** that is carried out after **enrollment** of specified numbers of **observation units** (usually persons), eg, for the first time when enrollment reaches 50 and again when it reaches 100, 150, etc. See DeMets and Ware [1980] and Pocock [1977]. ***Usage note***: See **interim data analysis**.

interim look *n* - [**trials**] 1. A **look** at the **results** of a trial while under way, especially when performed for the express purpose of determining whether the trial should be stopped or modified. 2. Any summary of **interim results** made during the course of a trial having the potential of being used to stop or modify the trial. 3. **interim result** rt: **multiple looks**, **administrative review**, **efficacy review**, **safety review**, **treatment effects monitoring**

multiple comparisons *n* - [**statistics**, in regard to two or more **comparisons** (defn 3) involving the same **dataset**] 1. Two or more comparisons involving the same **measure**; such comparison at the same point in time (as in the Coronary Drug Project[1973] involving comparison of **pairs** of treatments for a designated outcome at a single point in time); such comparison at different points in time (as in a particular **test-control treatment comparison** for a particular **outcome measure** at different points in time). 2. Two or more comparisons involving different measures; such comparisons at the same point in time (as in the Coronary Drug Project[1973] involving comparison of pairs of treatments for different outcomes at different points in time, or for **treatment group comparisons** for different **baseline characteristics**); such comparisons at different points in time (as in a particular **test-control treatment comparison** for different **outcome measures** at different points in time). 3. A comparison having an associated **p-value** or **confidence interval** that is **adjusted** to take account of the fact that it is one of several comparisons made or to be made. rt: **multiple looks**, **multiple outcomes** ***Usage note***: Virtually every **controlled trial** involves multiple comparisons in the sense of defns 1 and 2, even those involving just two **study treatments**. Any trial involving three or more study treatments and the need for two or more **pairwise comparisons** (as in the Coronary Drug Project[1973] in the comparison of each of 5 different test treatments with a **placebo control**) will involve multiple comparisons anytime the treatments are compared. Broadly inclusive of **multiple looks** and **multiple outcomes**.

multiple looks *n* **-** [**trials**] **Treatment comparisons** made at two or more time points over the course of a **trial**; especially when done in relation to **treatment effects monitoring** and where they may lead to alteration of the **treatment protocol**. rt: **multiple comparisons** ***Usage note***: Not to be confused with **multiple comparisons** as discussed in a usage note for that term.

multiple outcomes *n* **-** [**trials**] 1. The state or condition of having or being capable of yielding two or more **outcomes** (defn 1). 2. The state or condition of having or being capable of having two or more **outcome measures** for use in making **treatment comparisons**, as in a trial providing treatment comparisons for the **primary outcome measure** and for one or more **secondary outcome measures**. rt: **composite outcome**

premature stop *n* **-** [general] A **stop** occurring sooner than expected; one occurring prior to the planned finish or end. [**trials**] **premature termination** syn: **early stop**

primary outcome *n* **-** 1. [**trials**] The **event** or condition a **trial** is designed to **treat**, ameliorate, delay, or **prevent**. 2. The **outcome** of interest as specified in the **primary objective**. 3. The foremost measure of success or failure of a **treatment** in a trial. 4. The actual occurrence of a **primary event** in a **study participant**. 5. **primary endpoint** (not recommended; see usage note for **endpoint** for reasons). ***Usage note***: Not to be used interchangeably with **design variable**. The modifier, *primary*, should be used sparingly, since primariness depends on perspective. Most trials involve observations of various outcomes, each with different implications for well-being or life. There can be no doubt that death is a defining life event, but its importance as a life event does not mean that it is, therefore, always a good indicator of **treatment effect**. Sometimes, life can be worse than death, especially if life is reduced to a vegetative form.

primary outcome measure *n* **-** [**trials**] 1. That **measure**, among two or more, **observed** or to be observed in a trial that is considered to be of primary importance in its **design** (eg, the one used for the **sample size calculation**) or in the **analyses** performed or to be performed; may be a **continuous measure** or an **event** depending on the trial. syn: **primary outcome variable** 2. **design variable**

safety monitoring *v* **-** [**trials**] 1. **Monitoring** (defn 2) performed at periodic time points over the course of a trial, to determine whether the trial should be stopped or modified because of safety considerations; as distinct from **efficacy monitoring**. 2. **safety review** (defn 1) 3. **treatment effects monitoring** rt: **data and safety monitoring**, **efficacy monitoring**, **treatment effects monitoring**, **interim look**, **interim result** ***Usage note***: Often used in contradistinction to **efficacy monitoring** in settings where the user wishes to distinguish between **interim looks** performed for efficacy monitoring versus **safety monitoring**; eg, in settings where looks for safety monitoring are not counted as looks for purposes of adjusting **p-values** for **multiple looks**. The distinction is predicated on the assumption that safety and efficacy are **independent** dimensions of **treatment** — often not the case. Use **treatment**

effects monitoring when the distinction is unimportant or where the monitoring performed is for safety and efficacy. See also notes for **administrative review**, **efficacy monitoring** and **treatment effects monitoring**.

stopping guideline *n* **-** 1. A **guide** for determining when to stop or alter a **trial**. 2. A guide as to size or type of **treatment differences** that may cause **treatment effects monitors** to stop or alter a **trial**. rt: **stopping rule** ***Usage note***: Not to be used interchangeably with **stopping rule** (see for comments). Use **stopping guideline** instead of stopping rule if the rule is used simply as a guide as to when a stop or alteration may be indicated.

stopping rule *n* **-** 1. A **rule** for determining when to stop or alter a **trial**. 2. A rule for determining when to terminate or alter the **treatment protocol** of a **trial** based on the **observed treatment difference** for an **outcome** of interest; usually some function of a **p-value** produced by a designated **test statistic** evaluated at predetermined points in the course of the trial. The rule is an implicit part of the design in the case of **sequential trials**, it is established at or near the outset of the trial in the case of **fixed sample size designs**. A difference exceeding the set limit leads to **termination of trial** or one of the **study treatments**, depending on the nature and direction of the **observed treatment difference**. rt: **early stopping**, **stopping guideline** ***Usage note***: Not to be used interchangeably with **stopping guideline**. Reserve for uses where a stop or alteration proceeds more or less automatically when the conditions of the rule are met. Use **stopping guideline** if the rule is not binding.

Introduction

After a hiatus of months, I find myself again writing to you, this time on stopping rules. This memo has been brewing ever since my debate in the SOCA PDMB about two years ago regarding stopping rules.

My urge to put pen to paper was tweaked by a recent letter from the FDA regarding our IND for ADAPT because, in it, we asked, among other things, to explain why we have not specified stopping rules for ADAPT. But I guess that I would have been able to resist my penchant for letter writing had it not been for our first meeting of the ADAPT TEMC. It was that conference call that got me writing because, although no one on the call wants stopping rules, we are, nonetheless, debating whether we should have one so we can be "politically correct" in the world of clinical trials!

That stopping rules are desired in licensure trials is evident in the Statistical Principles for Clinical Trials in the Harmonised Tripartite Guidelines of the International Committee on Harmonisation (ICH). The section on interim analysis and early stopping specifies that: *The stopping guidelines and their properties should be clearly described in the protocol or amendments. ... If it becomes necessary to make changes to the trial, any consequent changes to the statistical procedures should be specified in an amendment to the protocol at the earliest opportunity, especially discussing the impact on any analysis and inferences that such changes may cause. The procedures selected should always ensure that the overall probability of type I error is controlled.*

The push for stopping rules by the FDA comes from an inherent distrust of sponsors (in the language of the FDA, *sponsor* is the holder of an IND, hence, we are the sponsor in ADAPT). They worry, that without rules, a sponsor will "dredge" to come up with something "significant".

One can appreciate why the FDA wants stopping rules because they want unassailable data if a trial is stopped because of benefit. But are imposed objectivity constructs a TEMC, such as stopping rules, good for patients? One can argue that they are not to the extent that that they keep a TEMC from acting in exercising its duty to preserve patients from harm.[1] One can argue that patients have the best protection by a highly competent TEMC not constrained by stopping rules or other imposed objectivity constructs.

On the difference between "stopping rule" and "stopping guideline"

Basically, a stopping rule is an algorithm, constructed prior to the start of the trial, that serves to indicate when a trial should stop. The use of a rule, under the frequentist construct, requires specification of the number of looks allowed during the trial and a "spending function" for the type I error. Most of the rule-based approaches to monitoring involve group sequential designs, eg, as represented by Lan-DeMets (1983), DeMets-Ware (1980), O'Brien-Fleming (1979), Pocock (1977) and Haybittle-Peto (1971).

Examples of stopping rules can be found in the HOPE (Heart Outcomes Prevention Evaluation) Study and in the CARET (Beta-Carotene and Retinol Efficacy Trial) [N Engl J Med 2000; 342:145-53; Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients; N Engl J Med 1996; 334:1,150-1,155; Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease]. HOPE: *Four formal interim analyses were planned. The statistical monitoring boundary indicating that ramipril had a beneficial effect was a difference in the primary outcome of 4 SD between groups during the first half of the study and of 3 SD during the second half. The respective boundaries indicating that ramipril had a harmful effect was 3 SD and 2 SD.* (The primary outcome in the trial was a composite event consisting of MI, stroke, or death from CV causes.) CARET: *The prespecified monitoring policy for stopping the trial early because of a benefit or adverse effect of the study vitamins was based on O'Brien-Fleming boundaries applied to the weighted number of confirmed lung-cancer end points, the primary end point. The critical P values were those of 0.0006 or lower for the first interim analysis in 1994 and those of 0.007 or lower for the second interim analysis in 1995.*

The expectation is that the trial will stop when a stop condition is reached and that the stop will occur as soon after the stop condition is reached as practical. The expectation is implicit in the name – <u>*stopping*</u> rule.

[1] Harm to persons in the context of trials can arise from continued use of a study treatment where accumulated data in the trial are sufficient to indicate that the treatment is harmful or from continued use of an ineffective treatment if the accumulated data are sufficient to indicate that a study treatment is superior to the one being administered.

A stopping guideline, as implied by the softer term, is merely a guide for when a stop may be indicated. Achieving the condition in the guideline does not mandate a stop.

The distinction may have operational meaning in the context of a trial, but it is lost on the broader community. In that setting, because stopping guidelines look like stopping rules, they are seen as stopping rules, even if called by a softer name. Hence, whatever the disadvantages for stopping rules, largely, the same disadvantages accrue to stopping guidelines.

On the use of stopping rules

The irony in regard to stopping rules is that they are rarely followed. That is, it is not uncommon for trials to be stopped before a stop is indicated for bad or nil effects or for them to be continued beyond the stop condition for good effects. An example of the latter is given by the HOPE Study. *On March 22, 1999, the monitoring board recommended termination of the study because of the clear evidence of a beneficial effect of ramipril (consistent crossing of the monitoring boundaries in two consecutive reviews).*

Assessing compliance to stopping rules is treacherous because of lack of documentation in publications. Even if a publication indicates that monitoring was done using a stopping rule, the publication is unlikely to indicate when the rule was devised or how often it was modified during the course of the trial.

Antariksha Kiri, for his PhD dissertation (2000) on treatment effects monitoring, reviewed a series of large trials ($n \geq 200$) published between 1990 and 1995. Of the 542 trials reviewed, 78 reported "early" stops. Among those, only 17 indicated that the stop was stopping-rule driven. Overall, only 44 of the 562 reports contained mention of preordained stopping rules for monitoring.

Reasons to distrust stopping rules

Reason 1: Because it is foolish to try to reduce a complex decision making process to a rule.

Comment

There are a myriad of conditions that can lead to early stops. It is impossible to anticipate all possible routes to early stops or to write viable rules when the trial is designed.

A stopping rule, by definition, is trial specific. Since it is to be written in advance of the trial it has to be based on a "guess" as to the underlying event rate for persons assigned to the control-treated group. Investigators are notorious for overestimating that rate and, hence, bad at writing realistic rules at the outset. The "fix" is to modify the rule later using data from the trial, but at the risk of investigators being seen as "fiddling" with the rule to make it do what investigators want.

A more basic problem is that they are written as if the only information relevant to stopping is internal to the trial. But life is rarely so simple, especially in long-term trials. Reports from other trials during the trial can impact on decision making in the trial. A case in point is

CARET. The largely negative results from the Alpha-Tocopherol Beta Carotene Cancer Prevention Study published in 1994 [N Engl J Med 1994; 330:1,029-1,035] contributed to an early stop of CARET because of lack of effect and possible bad effects in CARET.

Reason 2: Because rule-driven decision making is at odds with the competency requirement in monitoring.

Comment

See *IRBs and randomized clinical trials* in IRB, vol 20, March-June, pgs 9-12 for discussion of requirements for competency in monitoring.

Stopping rules are objectivity constructs imposed on TEMCs to protect against early stops. They exist to protect the scientific integrity of the trial. The protection is the expense of monitoring competency and, hence, has potential for risks to persons studied.

Reason 3: Because the purpose of a rule is to restrict the number of looks that can be made.

Comment

Restricting the number of looks is at odds with the competency requirement for monitoring. One can argue that the goal in creating TEMCs should be to recruit the most qualified and experienced people possible and then to give them free reign to act as they see fit. Imposing restrictions on the number of looks that can be made and in what can be looked at are gimmicks imposed to deal with flaws in the frequentist approach to statistical inference. It is difficult to argue that such restrictions increase protections for persons studied.

Reason 4: Because the restriction on the number of looks is illusory.

Comment

A good coordinating center looks at treatment differences repeatedly over the course of trial. Indeed, one can argue that it has a moral duty to do so and to alert study officers and the TEMC to differences suggestive of harm, even if the alert does not arise in relation to a "scheduled" look.

The "fix" to making the restriction less illusionary is to enjoin the coordinating center from looking, except in relation to scheduled looks. However, that "fix" raises ethical concerns in that it serves to reduce the monitoring competency of the center and is at odds with its duty to preserve persons from harm.

So when is a look a look? Is the TEMC charged with a look when it meets? Do conference calls count in which data are discussed count as "looks"? Do all looks count or are some "unofficial"? These are the questions that move to the fore when TEMCs are faced with restrictions on the number of look allowed.

Reason 5: Because the rule is not a rule.

Comment

Most trials stop short of the rule if the trend is in the wrong direction and proceed beyond the stop condition if the trend is the right direction. For example, CARET stopped well short of its stop condition for bad effects and the HOPE Study kept going beyond the stop condition for good effects. There are good reasons for both types of behavior, but why bother writing rules if they are not likely to be followed?

Reason 6: Because a rule can turn into a club.
Comment
The conversion happens when the results are published. Early stops are likely to be controversial, especially if the stop was for bad effects and the treatment is in widespread use, eg, as in the case of the stop for tolbutamide in the UGDP. You are likely to receive incoming fire regardless of how you decided to stop. So why give your critics added ammunition by a "rule" you did not follow?

The reality is that a rule, once written, is like Freddie Krieger, of Friday the 13th fame – impossible to kill and impossible to bury. If it existed, even if subsequently disavowed by the study investigators, someone will dig it up and ask you to explain why it was not followed. Even if you follow the rule there will be those who say the rule was wrong and that your slavish adherence to it caused harm to patients.

Reason 7: Because rules tend to substitute for creative thought.
Comment
The tendency, with a rule, is for the TEMC to focus on the rule. In effect, the rule becomes the issue. Instead of spending its energy exploring data, the TEMC spends its time arguing about the rule and when a look is a look.

Reason 8: Because the use of a stopping rule implies that monitoring is merely statistical in nature.
Comment
The rule, because it is statistical in nature, implies that it is the "statisticians" who make the call. But statisticians, individually or collectively, are not that smart or all knowing. There is a statistical element in monitoring and decision making to be sure, but it is only one element.

Reason 9: Because of the tendency to write rules to preclude early stops for ill-effects.
Comment
Evidence of this tendency is seen for both the HOPE Study and CARET. The lower bound for both studies, and especially for CARET, imply a willingness to continue to "proof" of harm. One can raise serious ethical questions about such postures.

Reason 10: Because the rule must be symmetrical about the no difference line if the TEMC is masked.
Comment
Though I have written on why masked monitoring is ill-advised, it remains fairly widely practiced. Under masking, the boundaries have to be symmetrical about the no difference line because the TEMC does not know if the difference favors the test or the control treatment. But, symmetrical boundaries imply that the TEMC is as interested in "proving" that the treatment is bad as it is in "proving" it is good. Continuing a trial to "prove" harm is ethically untenable.

The reality is that trials stop far short of "proof" in the face of negative results, as they should. Hence, it is obvious that the lower bound for a rule, such as in CARET, is artifactual. (CARET had symmetrical boundaries because the TEMC was masked to treatment group.) The "fix" in CARET is telegraphed in the paper: *When the results of the ATBC Cancer Prevention Trial became available, the committee reviewed the results of the first interim analysis and requested that the blinding be ended. Subsequently, the committee reviewed data unblinded.*

Reason 11: Because the rule is focused on efficacy.
Comment
As already suggested in comments for Reasons 9 and 10, the lower bound for most rules are likely to be ignored unless they are nearer the no difference line than is the case for the HOPE Study or for CARET. As a result, the tendency is to construct rules relating only to efficacy. The shortcoming in that approach is that the focus implies that efficacy can be assessed independently of safety, whereas the reality is it cannot be. Efficacy has to evaluated against safety.

One of the reasons for the focus is that stopping rules require restrictions on the number of looks allowed. That restriction is ill-advised when it comes to issues of safety. The "fix" in regard to the restriction is to separate the two monitoring processes so that "looks" for safety do not count as looks for efficacy. The separation is artificial.

A parenthetical note in passing: It is ironic that the usual name for monitoring committees is *Data and safety monitoring committee* – not *Data and efficacy monitoring committee*.

Reason 12: Because the rule is limited to one outcome measure.
Comment
All trials are multivariate in nature. Typically, the sample size is rationalized using a particular outcome measure, but persons are observed for a variety of outcomes. However, stopping rules, of practical necessity, are designed to focus on just one outcome measure – the design variable or primary outcome measure. This focus leads to the mindless view that the only variable investigators can talk about or draw conclusion from is that variable. Why? Because it was the one specified in the stopping rule, none of the other measures are "eligible" for inference.

So if the design variable was occurrence of MI, does that mean that one would ignore differences in mortality? One would hope not. Yet, from the stopping rule perspective, the only difference that matters is the one for the outcome measure specified in the stopping rule.

The notion that the only things that can be looked at in trials are those outcome measures specified before the trial started is foolish. It leads to absurdities of the kind raised by critics of the UGDP, when the first results were published in 1970. Critics of the decision to stop use of tolbutamide argued that the difference in mortality should have been ignored because investigators had not specified that they would look at mortality in the protocol!