

Approaches to Monitoring the Results of Long-Term Disease Prevention Trials: Examples from the Women's Health Initiative

Laurence Freedman, Garnet Anderson, Victor Kipnis, Ross Prentice, C.Y. Wang, Jacques Rossouw, Janet Wittes, and David DeMets

National Cancer Institute, Bethesda, MD (L.F., V.K.), Fred Hutchinson Cancer Research Center, Seattle, WA (G.A., R.P., C.Y.W.), National Institutes of Health, Bethesda, MD (J.R.), Statistics Collaborative, Washington, DC (J.W.), and University of Wisconsin, Madison, WI (D.D.)

ABSTRACT: We contrast monitoring therapeutic trials with monitoring prevention trials. We argue that in monitoring prevention trials one should place more emphasis on formally defined global measures of health, not simply on a single targeted disease, particularly when an intervention may reduce the incidence of some diseases but increase the incidence of others. We describe one approach, illustrated by the Women's Health Initiative. For each of several sets of hypothetical interim results ("scenarios"), members of the Data and Safety Monitoring Committee (DSMC) were asked whether they would continue or stop the trial. In parallel with this exercise, various statistical methods of monitoring that are based on (1) the primary targeted disease, (2) a combination of various disease outcomes, or (3) a mixture of both were applied to these scenarios. One objective was to find a statistical approach that mirrors the majority view of the DSMC. A second objective was to stimulate discussion among DSMC members in preparation for their task of monitoring the trial as the real data become available. We found that no single method fully matched the majority vote of the DSMC. However, a mixed approach requiring the primary outcome to be significant and the global index to be "supportive," with separate monitoring of adverse effects, corresponded with the majority vote quite well. This approach maintains the emphasis on the primary hypothesis while assuring that broader safety and ethical issues of multiple diseases are incorporated. © Elsevier Science Inc., 1996 *Controlled Clin Trials* 1996; 17:509-525

KEY WORDS: *Clinical trials, disease prevention, data and safety monitoring, multiple endpoints, stopping rules*

INTRODUCTION

It is now increasingly accepted that clinical trials need careful monitoring to ensure the safety of the participants and the ethics of continued experimentation [1, 2]. For long-term trials with many participants, it is also becoming the

Address reprint requests to: Laurence Freedman, Biometry Branch, DCPC, Division of Cancer Prevention and Control, National Cancer Institute, Executive Plaza North, Suite 344, 6130 Executive Boulevard MSC 7354, Bethesda, MD 20892-7354.

Received August 8, 1995; revised January 22, 1996; accepted February 7, 1996.

norm to establish a group of experts, independent of the trial investigators, to perform this monitoring task. In this paper, we call such a group a Data and Safety Monitoring Committee (DSMC).

One of the most difficult issues that a DSMC faces is determining how much evidence in support of the superiority of one of the treatments can be allowed to accrue before a trial is stopped. To answer this question, the DSMC must balance scientific, statistical, and ethical considerations, often of considerable subtlety [3]. Therefore, the typical DSMC includes experts from the relevant clinical and basic science disciplines, statisticians, and ethicists. To aid the DSMC in its weighing of evidence, statisticians have developed a considerable body of theoretical and applied statistical methodology on "stopping rules." Most of this work is devoted to specifying the level of statistical significance that a treatment difference should attain before the DSMC recommends terminating a trial. Usually the methodology assumes that one disease outcome or measure is of primary importance, such as time of survival from entry to the trial or time to recurrence of disease. The stopping rule is then based on a comparison of the treatment groups with respect to that measure.

In this paper we argue that a single outcome is often not an appropriate paradigm for trials of disease prevention. Our remarks mainly address primary prevention trials, i.e., trials involving interventions given to ostensibly healthy individuals to reduce their risk of developing a certain disease. We propose a new approach to developing statistical stopping rules for such trials. We envisage that this approach would be applicable to several current long-term prevention trials, such as the Breast Cancer Prevention Trial [4] and the Beta-Carotene and Retinol Efficacy Trial (CARET) [5]. In this paper, our motivating example is the Women's Health Initiative (WHI) Clinical Trial. Our description of the trial design and the disease outcomes of interest follow the initial protocol. *There was a major revision in the design of the Hormone Replacement Therapy component in January 1995 (described in the footnote to Figure 1), that is not reflected in this paper. However, since we discuss a general method, with WHI as an illustration, the assumptions and conclusions of this paper are not affected.*

The WHI Clinical Trial has three components: dietary modification, hormone replacement therapy, and calcium/vitamin d supplementation [6]. Each component involves a separate randomization (see Figure 1). Each component has one or more "primary" diseases of particular interest because preliminary evidence regarding the effects of the intervention on them is quite extensive and because these hypothesized effects motivated the trial. In addition, each component includes other "secondary" diseases for which the preliminary evidence is somewhat less extensive, or that have somewhat less serious health consequences, or that may represent possible hazardous consequences of the intervention. Table 1 lists the primary and secondary diseases. The protocol of the WHI trial reviews the evidence for the effect of the intervention on each of these diseases [6].

MONITORING PREVENTION TRIALS AND THERAPEUTIC TRIALS

Treatment and prevention trials have some important commonalities. Both types of trial are typically designed to answer a primary question relating to the effect of an intervention on a specific disease. These questions, which are

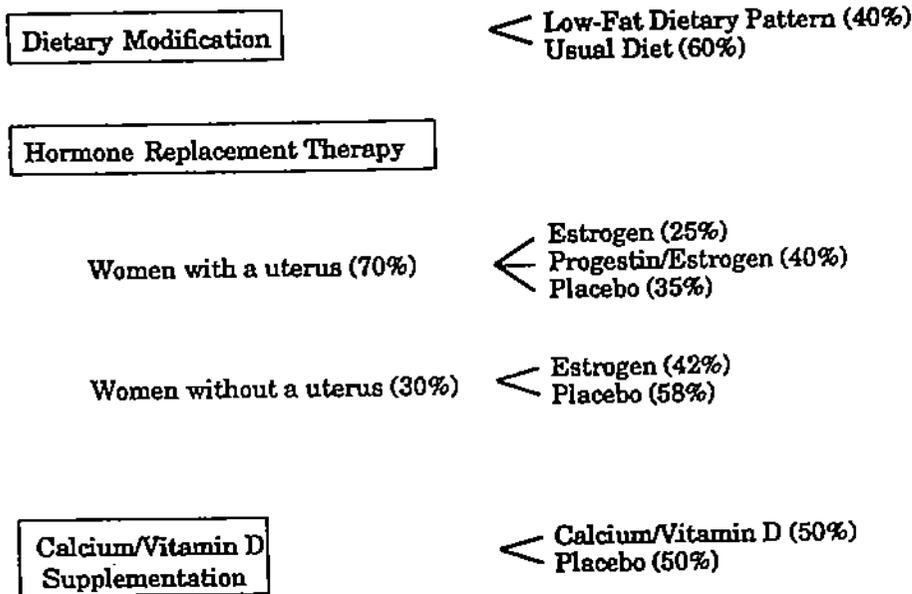


Figure 1 Design of the Women's Health Initiative Clinical Trial according to the initial protocol. The design of the hormone replacement therapy component was changed in January 1995. The principle change was to drop the unopposed estrogen option for women with a uterus. The sample size was increased to 27,500 from 25,000, and the proportion of women without a uterus was increased from 30% to 45%. Trial participants are required to enter either the dietary randomization or the HRT randomization, and may opt for both. In addition they may enter the calcium/vitamin D randomization 1 year later. Figures in parentheses are the proportions of participants randomized to a given intervention in that part of the trial.

based on hypotheses generated by previous study results, provide the rationale for the trial. The conventional approach to monitoring a trial focuses strongly on the specific hypothesis or disease. We think that this perspective derives from and is appropriate for many treatment trials. However, we believe that the strong focus on one specific disease is often inappropriate for prevention trials. The following are some reasons why methods for monitoring disease prevention trials may have to be different from those used for monitoring treatment trials.

1. Subjects entering treatment trials suffer from a defined disease or condition and are seeking alleviation of its consequences. The main monitoring index can therefore be sensibly confined to some aspect of such consequences, e.g., time to disease recurrence or to death from the disease. When most deaths are expected to be caused by the targeted disease, time to death from any cause is sometimes chosen as the main index. In contrast, subjects entering disease prevention trials are ostensibly healthy. Although the intervention is targeted to affect certain diseases, these diseases may constitute only a small part of the health concerns of a typical participant in the trial.

Table 1 Disease Components of the Combined Index and Their Weights for Each of the Randomizations in the WHI

	Dietary Modification	Hormone Replacement Therapy	Calcium/Vitamin D
Breast cancer incidence*	Weight 0.35	Weight 0.50	Weight 0.18
Colorectal cancer incidence*	0.50	0.18	0.50
CHD incidence	0.50	Breast cancer incidence	Hip fracture incidence*
Deaths from other causes	1.00	Endometrial cancer incidence	Colorectal cancer incidence
		Deaths from other causes	Deaths from other causes
		1.00	1.00

*Primary endpoints.

2. In most treatment trials, there is considerable morbidity or mortality within a few years of the beginning of the trial. The focus on the effect of treatment is therefore over a relatively short term. The finding that a treatment reduces mortality or disease recurrence over the first few years is often judged sufficient to offset the chance of discovering later deleterious effects. Even when such effects are discovered they are often tolerated for the sake of the early benefits. For example, cytotoxic chemotherapies are used for treating cancers, even though some agents can cause the development of new malignancies later in life. In contrast, in prevention trials morbidity and mortality rates are typically low. Even where an intervention is shown to reduce the incidence of a "common" disease by half, the benefit may accrue to only a few percent of the participants. Moreover, the morbidity or mortality from the targeted disease will usually represent only a small proportion of the total morbidity or mortality experienced by the participants. Consequently, prevention trials have much greater potential for unexpected effects of the intervention to overshadow the expected beneficial effects and for later effects of the intervention to overshadow earlier effects.
3. The interventions in prevention trials often carry potential effects on *several* diseases. These effects may be beneficial for some diseases and adverse for others. Furthermore, there may already exist strong evidence for some of these effects but weak evidence for others. Thus, a major rationale for the trial may be to estimate all of these effects more precisely, thereby providing the basis for informed public health decisions. DSMCs therefore need methods for comparing and balancing these benefits and risks to help guide the monitoring of trials. To further complicate the task of monitoring a prevention trial, the time course of potential beneficial and adverse effects may differ considerably.
4. Because treatment trials are usually smaller and of shorter duration than prevention trials, they are more likely to be repeated if their results are equivocal. Generally, a large-scale prevention trial is unlikely to be repeated because of its long-term nature and its expense. The results of prevention trials are thus more likely to be translated directly into practice or even public health policy. Therefore it is necessary to ensure that the effects of interventions in prevention trials are assessed as thoroughly and broadly as possible, and that, in the absence of adverse effects that would eliminate the public health utility of an intervention, the trial is not stopped until a clear answer is obtained to the broad question: "Does this intervention give an overall benefit in health to the population?"

As a consequence of these differences between treatment and prevention trials, we argue for a more comprehensive approach to monitoring of prevention trials. Specifically, we propose that, in addition to considering the effect of the intervention on primary outcomes, overall health benefit vs. risk considerations be incorporated into formal stopping procedures. In other words, we wish to balance in the monitoring process the requirement for *global* assessment of health effects with the requirement to confirm or deny certain hypotheses about the effect of the intervention on *specific* diseases. In the next section we will suggest some different approaches to monitoring prevention trials that vary

from a purely "global" approach to a purely "specific" approach, along with combinations thereof.

METHODS FOR THE STATISTICAL MONITORING OF PREVENTION TRIALS

Unless otherwise specified, we envisage using a group sequential method [2], e.g., the O'Brien and Fleming method [7], on a defined outcome measure. In the group sequential method one calculates, at each interim analysis, the difference in the occurrences of the defined outcome in the two groups and applies a statistical test of the null hypothesis of no difference. For example, one may use the binomial distribution to test a difference in the proportions of disease occurrences. The critical value of the test statistic depends on the planned number of analyses, the order of the present analysis in the planned sequence, and the choice of group sequential method [8]. For the WHI Clinical Trial, three analyses are planned. Using the O'Brien and Fleming method, the critical values of the standardized test statistic, z , for the three analyses are 3.47, 2.45, and 2.00, respectively, which preserve an overall two-sided significance level of 5%.

The following discussion introduces some possible outcome measures or monitoring indices. The indices that we describe are intended for monitoring the trial for overall health benefit. At the end of the section, under "mixed methods," we suggest how to incorporate monitoring for adverse effects.

Purely Global Approaches

The simplest purely global outcome measure and one that has been used previously in cardiology trials, *total mortality*, has many advantages. It is easy to ascertain, calculate, and interpret. Demonstrating a significant sustained reduction in the total mortality of the intervention group provides strong argument for the termination of the trial and subsequent use of the intervention. The global nature of the measure assures that the assessment covers all potentially fatal diseases.

However, the measure also carries certain disadvantages. First, total mortality may be somewhat insensitive to strong effects on one or more serious diseases because (1) it includes only deaths and not incident cases of these diseases, thus counting fewer disease events; (2) these effects will be diluted by inclusion of deaths from other diseases that are unaffected by the intervention; and (3) intervention effects on morbidity may be detected some years before corresponding mortality differences are evident. Second, the DSMC may consider it unsatisfactory to stop the trial on the basis of a reduction in total mortality, in the absence of demonstrating the effect of the intervention on one or more specific diseases. The result might be regarded as purely statistical without any clear biological explanation. Furthermore, investigators and study participants may be uneasy about stopping the trial before it provides a clear answer to the hypotheses specified in the protocol.

An alternative global measure based on *total morbidity* instead of total mortality is infeasible because of the intractable problems of its definition and ascertainment. We will not consider total morbidity among our potential monitoring indices, but we mention it here for completeness.

Purely Specific Approaches

Conventional methods of monitoring trials use the *primary endpoint* as the outcome measure. Table 1 lists the primary endpoints for each component of the WHI Clinical Trial. This measure has the advantage of being closely linked to the sample size calculations and design parameters on which the trial is based. In addition, it should be very sensitive to the effect specified to be of primary interest. However, its use may promote (even if it does not necessitate) undue focus on the specific diseases. In addition, in situations where reductions in some diseases and increases in other diseases occur (e.g., with hormone replacement therapy there is an anticipated benefit to coronary heart disease but potential increased risks of breast and endometrial cancer), this measure gives little or no help to the decision making process. Members of the DSMC are left to weigh benefits and risks in an ad hoc manner, possibly in an environment of intense external and internal pressure. Arriving at sensible decisions under such circumstances is never guaranteed, but the availability of some objective procedures is likely to be helpful.

An alternative specific approach is to use *mortality from the disease of primary interest*, or cause-specific mortality, instead of incidence. This mortality measure retains the disadvantages but not the advantages of the incidence measure (assuming that the trial is designed with incidence as the primary outcome) and we do not consider it further.

Combined Index Approaches

We consider some indices that combine effects of several diseases specified in the protocol as of special interest together with a global effect on "other diseases." For example, in the dietary modification component of the Women's Health Initiative, the diseases of special interest are breast cancer, colorectal cancer, and coronary heart disease. One possible index combines the proportions diagnosed with each of these three diseases, together with the proportion dying from other causes. We propose three variants of the combined index: (1) an unweighted combined index; (2) a weighted combined index; and (3) a weighted combined index with Bayesian (as opposed to group sequential) monitoring.

Unweighted Combined Index

Let $d_1, d_2, d_3, \dots, d_k$ be the observed differences in proportions for the outcomes 1, 2, . . . , k (e.g., as in Table 1 for the WHI Clinical Trial). The unweighted index, U , is defined by:

$$U = d_1 + d_2 + d_3 + \dots + d_k \quad [1]$$

The statistic U is monitored the same way as a primary outcome difference, using O'Brien and Fleming boundaries. The unweighted index simply counts disease events. Note, however, that the index combines the number of *diagnoses* of the diseases of special interest plus the number of *deaths* from other causes. Thus, the index emphasizes the diseases of special interest but does not ignore other fatal diseases. One can think of this index as a type of total mortality measure that replaces deaths from the diseases of special interest by diagnoses

of these diseases. Since these diseases are not invariably fatal, the index thereby inflates their contribution. This measure, attractive in its simplicity, offers a reasonable compromise in balancing the issue of generality and specificity mentioned in our rationale. A simple version of the measure, combining nonfatal myocardial infarction with total mortality events, has been used in some cardiology trials.

Weighted Combined Index

An objection to the unweighted index is that in some contexts the specific diseases contained in the index may be of very different severity, yet the index counts all diagnoses as equal. To accommodate this consideration we propose weighting the occurrence of each disease by the expected proportion of diagnosed persons who will die of that disease within a specific number of years of diagnosis. These weights are necessarily ≤ 1.0 . The weight for deaths from other causes will remain as 1.0. Let w_1, w_2, \dots, w_k represent these weights for the outcomes 1, 2, \dots , k . They may be obtained from data external to the trial. The weighted index W is defined by

$$W = w_1d_1 + w_2d_2 + w_3d_3 + \dots + w_kd_k \quad [2]$$

For the WHI Clinical Trial we chose the period following diagnosis to be 10 years. The value of w for each WHI outcome is specified in Table 1. The statistic W is monitored the same way as U .

This weighting has two consequences. First, the diseases are weighted according to their likelihood of resultant deaths. Second, the index becomes a measure of predicted total mortality. As in the unweighted index, the special interest diseases receive more emphasis than other diseases, but now the emphasis results from counting for the specific diseases the deaths that are predicted to occur over the succeeding specified period but counting deaths from other diseases only as they occur. The rationale for the weighted index is similar to that for the unweighted index. The weighted index is closer to the "total mortality" end of the scale, with somewhat less emphasis on the specific diseases than the unweighted index.

Weighted Combined Index with Bayesian Monitoring

Another consideration in the use of such combined indices is the large variation in the strength of preliminary evidence for an intervention effect on each of the specified diseases and on deaths from other causes. For example, in the Dietary Modification component of WHI, there is less preliminary evidence for a reduction in deaths from other causes than for a reduction in breast cancer, colorectal cancer, or coronary heart disease. To accommodate this, we could use Bayesian methods based on the weighted combined index, with skeptical prior distributions [9]. These methods effectively introduce a further weighting of each endpoint according to the level of preliminary evidence concerning the effect of the intervention on that endpoint. The greater the evidence, the larger the weight placed on that endpoint. The level of evidence would usually be least for deaths from other causes. This approach will tend to move the index away from the "total mortality" end of the scale toward the "specific diseases" end.

Mixed Approaches

Until now, we have discussed the use of single outcome measures that can be monitored using standard statistical methods for monitoring clinical trials. Possibly none of these single measures will give entirely satisfactory results, whereas a mixture of methods may perform better. Two possible types of mixture for the assessment of beneficial effects are:

1. Requiring that both the primary endpoint and a more global measure reach significance when monitored as separate single outcome measures. We could mix the primary endpoint with either total mortality or any of the three types of combined index described above. For example, in the WHI Clinical Trial, at the second of three analyses, we would require that both the primary outcome and the chosen global measure be significant at $|z| > 2.45$, the O'Brien and Fleming critical z value.
2. Requiring that the primary endpoint reach significance and that a more global measure (total mortality or a combined index) be "supportive" of the primary endpoint result. For example, significance at a nominal level of 20%, rather than the customary 5% level, could be required. For the WHI Clinical Trial, at the second analysis, we would require that the primary outcome be significant at $|z| > 2.45$ and that the global measure be significant at $|z| > 1.69$.

The main rationale for such mixtures is recognition that each aspect of the results, both global and specific, could be viewed as having importance in terms of the recommendations to be given to the targeted population and in terms of their likely adoption; the mixture approach allows the specific disease results to "veto" stopping the trial on the basis of the global results, and vice versa.

Mixture approaches may be useful in monitoring prevention trials not only for beneficial effects but for adverse effects. For example, in the hormone replacement therapy component in the WHI, one could separately monitor breast cancer as a potential adverse effect. Then, in a mixture approach, a recommendation to terminate would be made *either* if a formal stopping criterion based on the adverse effect were reached, *or* if one of the stopping criteria for benefit, such as (2) above, were reached. For the WHI Clinical Trial, at the second analysis, we would require either that the adverse effect be significant at $|z| > 2.45$, or that the primary outcome be significant at $|z| > 2.45$ and the global measure be significant at $|z| > 1.69$.

We do not advocate stopping criteria for adverse effects to be formulated as requiring both the adverse effect and the global measure to reach their corresponding significance levels. We believe that the potential benefits and risks are not symmetrical issues in prevention trials and that practically important safety issues must take precedence.

SCENARIOS FOR CHOOSING A STATISTICAL STOPPING RULE

In the last section we described several statistical methods of monitoring a prevention trial and advising on early termination. In the complex situations that are envisaged it would be unwise to adopt any one of these without

Table 2 Scenario 1—Diet: Incidence and Mortality Rates Correspond to Those Assumed in the Trial Design

	6 Years		z
	C (n = 28800)	I (n = 19200)	
Incidence ^a			
Breast Cancer	2.05 (0.08)	1.85 (0.10)	1.56
Colorectal Cancer	1.07 (0.06)	0.92 (0.07)	1.63
CHD	3.02 (0.10)	2.63 (0.12)	2.54*
Mortality ^b			
Breast Cancer	0.51 (0.04)	0.46 (0.05)	0.78
Colorectal Cancer	0.37 (0.04)	0.32 (0.04)	0.97
CHD	1.21 (0.06)	1.05 (0.07)	1.64
Other causes	5.50 (0.13)	5.50 (0.16)	0.00

^aThe percentage of incident cases in each group (standard errors in brackets).

^bThe percentage of deaths in each group (standard errors in brackets).

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming.

ensuring that members of the DSMC are comfortable with its use. Moreover, we would want a method that is consistent with good clinical judgment. For the purpose of identifying such a method, a series of hypothetical sets of interim results (scenarios) for the trial were formulated. Each member of the DSMC was asked to consider the scenarios and recommend whether to stop the trial. In parallel with this exercise several of the statistical methods described in the previous section were applied to the same scenarios.

This exercise can be expected to yield several lessons. First, examining the level of disagreement among the DSMC members on each scenario will identify in advance situations that are likely to cause some controversy within the committee over the appropriate course of action. Second, comparing the results of the statistical methods with the opinions of the DSMC will indicate which statistical methods, if any, may be acceptable to the members. Third, as an important byproduct, the DSMC can use the scenarios to discuss broad issues underlying the clinical, ethical, and statistical aspects of monitoring the trial. It is not necessary that one or more of the scenarios occur during the trial for the exercise to be useful. Discussion of the scenarios can serve as a valuable training exercise for the DSMC, in preparation for the real-life experience of monitoring the trial.

To apply this approach to the Women's Health Initiative Clinical Trial, we constructed eight scenarios: three for the dietary modification component, four for hormone replacement, and one for calcium/vitamin D supplementation (Tables 2–9). Each scenario comprises hypothetical results after an average 6 years follow-up; the trial is designed for an average 9 years follow-up. For each scenario we specified the incidence and mortality rates of the primary and secondary diseases in the intervention and control groups and the mortality rates due to other causes. We also provided the z value of the group comparison for each disease outcome.

The scenarios were aimed at exploring potentially difficult situations, e.g., where evidence of beneficial effects on a secondary disease appears strong

Table 3 Scenario 2—Diet: Incidence and Mortality Rates as Used in Trial Design Except that Mortality from Other Causes Is Reduced by 7% in the Dietary Modification Group

	6 Years		z
	C (n = 28800)	I (n = 19200)	
Incidence^a			
Breast Cancer	2.05 (0.08)	1.85 (0.10)	1.56
Colorectal Cancer	1.07 (0.06)	0.92 (0.07)	1.63
CHD	3.02 (0.10)	2.63 (0.12)	2.54*
Mortality^b			
Breast Cancer	0.51 (0.04)	0.46 (0.05)	0.78
Colorectal Cancer	0.37 (0.04)	0.32 (0.04)	0.97
CHD	1.21 (0.06)	1.05 (0.07)	1.64
Other causes	5.50 (0.13)	5.11 (0.16)	1.85

^aThe percentage of incident cases in each group (standard errors in brackets).

^bThe percentage of deaths in each group (standard errors in brackets).

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming.

but evidence on the primary disease does not, or where there are apparently beneficial effects on some diseases and harmful effects on others. As the results show, this policy did indeed lead to our identifying situations in which the DSMC opinion was seriously split.

Table 10 shows the voting of individual members for each scenario. The scenarios (numbers 4 and 7) causing the most disagreement among members were those involving hormone replacement therapy in which an excess of endometrial cancer is observed. In scenario 4 the DSMC disagreed as to whether the beneficial effects observed on coronary heart disease and osteoporosis are sufficient to offset the increase in endometrial cancer incidence and terminate

Table 4 Scenario 3—Diet: Incidence of Breast and Colorectal Cancer Are Reduced Significantly in the Dietary Modification Group; No Change in CHD or Deaths from Other Causes

	6 Years		z
	C (n = 28800)	I (n = 19200)	
Incidence^a			
Breast Cancer	2.05 (0.08)	1.72 (0.09)	2.63*
Colorectal Cancer	1.07 (0.06)	0.83 (0.07)	2.69*
CHD	3.02 (0.10)	3.02 (0.12)	0.00
Mortality^b			
Breast Cancer	0.51 (0.04)	0.43 (0.05)	1.27
Colorectal Cancer	0.37 (0.04)	0.29 (0.04)	1.59
CHD	1.21 (0.06)	1.21 (0.08)	0.00
Other causes	5.50 (0.13)	5.50 (0.16)	0.00

^aThe percentage of incident cases in each group (standard errors in brackets).

^bThe percentage of deaths in each group (standard errors in brackets).

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming.

Table 5 Scenario 4—Hormone Replacement Therapy†: Unopposed Estrogens Reduce Incidence of CHD and Hip Fractures Significantly; Breast Cancer Incidence Is Increased 20% in the ERT Group; Endometrial Cancer Increased 4-fold; No Change in Deaths from Other Causes

	6 Years		z
	C (n = 10500)	I (n = 7500)	
Incidence^a			
CHD	3.26 (0.17)	2.59 (0.18)	2.66*
Hip fractures	1.87 (0.13)	1.37 (0.13)	2.65*
Breast Cancer	2.07 (0.14)	2.25 (0.17)	-0.82
Endometrial Cancer	0.46 (0.07)	1.30 (0.13)	-5.72*
Mortality^b			
CHD	1.30 (0.11)	1.04 (0.12)	1.61
Hip fractures	0.47 (0.07)	0.34 (0.07)	1.37
Breast Cancer	0.52 (0.07)	0.56 (0.09)	-0.36
Endometrial Cancer	0.05 (0.02)	0.13 (0.04)	-1.80
Other causes	5.37 (0.22)	5.37 (0.26)	0.00

† This scenario is based on the initial protocol. There was a major revision in the design of the HRT component in January 1995 that is not reflected here.

^a The percentage of incident cases in each group (standard errors in brackets).

^b The percentage of deaths in each group (standard errors in brackets).

* Exceeds the 5% critical level of 2.45 using O'Brien and Fleming.

Table 6 Scenario 5—Hormone Replacement Therapy†: Progestin/Estrogen Reduces Incidences of CHD and Hip Fractures as Designed; Increases Incidence of Breast Cancer 20%; No Change in Deaths from Other Causes

	6 Years		z
	C (n = 6125)	I (n = 7000)	
Incidence^a			
CHD	3.26 (0.23)	2.60 (0.19)	2.23
Hip fractures	1.87 (0.17)	1.49 (0.14)	1.68
Breast Cancer	2.07 (0.18)	2.25 (0.18)	-0.71
Endometrial Cancer	0.46 (0.09)	0.46 (0.08)	0.00
Mortality^b			
CHD	1.30 (0.14)	1.04 (0.12)	1.38
Hip fractures	0.47 (0.09)	0.37 (0.07)	0.88
Breast Cancer	0.52 (0.09)	0.56 (0.09)	-0.31
Endometrial Cancer	0.05 (0.03)	0.05 (0.03)	0.00
Other causes	5.37 (0.29)	5.37 (0.28)	0.00

† This scenario is based on the initial protocol. There was a major revision in the design of the HRT component in January 1995 that is not reflected here.

^a The percentage of incident cases in each group (standard errors in brackets).

^b The percentage of deaths in each group (standard errors in brackets).

Table 7 Scenario 6—Calcium/Vitamin D: Incidence of Hip Fractures and Colorectal Cancer Is Reduced as Designed: 1% Reduction in Deaths from Other Causes

	5 Years		z
	C (n = 22500)	I (n = 22500)	
Incidence ^a			
Hip fractures	1.51 (0.08)	1.21 (0.07)	2.75*
Colorectal Cancer	0.86 (0.06)	0.75 (0.06)	1.31
Mortality ^b			
Hip fractures	0.38 (0.04)	0.30 (0.04)	1.46
Colorectal Cancer	0.30 (0.04)	0.26 (0.03)	1.02
Other causes	5.92 (0.16)	5.86 (0.16)	0.27

^aThe percentage of incident cases in each group (standard errors in brackets).

^bThe percentage of deaths in each group (standard errors in brackets).

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming.

the trial with a recommendation to use estrogen replacement therapy. In scenario 7 they disagreed over whether the nonsignificant benefits on coronary heart disease and osteoporosis are so small as to close the trial and advise against using estrogen without progestin. (Following the January 1995 revision of the WHI protocol, neither of these scenarios is relevant, because women with a uterus are no longer randomized to estrogen without progestin.) Scenarios 3 and 6

Table 8 Scenario 7—Hormone Replacement Therapy†: Unopposed Estrogen Reduces Incidence of CHD and Hip Fractures by Only 1/3 of the Designed Effect; Increases Incidence of Breast Cancer 40%; Increases Incidence of Endometrial Cancer Four-fold; No Change in Deaths from Other Causes

	6 Years		z
	C (n = 10500)	I (n = 7500)	
Incidence ^a			
CHD	3.26 (0.17)	3.04 (0.20)	0.84
Hip fractures	1.87 (0.13)	1.74 (0.15)	0.65
Breast Cancer	2.07 (0.14)	2.43 (0.18)	-1.60
Endometrial Cancer	0.46 (0.07)	1.30 (0.13)	-5.72*
Mortality ^b			
CHD	1.30 (0.11)	1.22 (0.13)	0.48
Hip fractures	0.47 (0.07)	0.44 (0.08)	0.30
Breast Cancer	0.52 (0.07)	0.61 (0.09)	-0.79
Endometrial Cancer	0.05 (0.02)	0.13 (0.04)	-1.80
Other causes	5.37 (0.22)	5.37 (0.26)	0.00

†This scenario is based on the initial protocol. There was a major revision in the design of the HRT component in January 1995 that is not reflected here.

^aThe percentage of incident cases in each group (standard errors in brackets).

^bThe percentage of deaths in each group (standard errors in brackets).

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming.

Table 9 Scenario 8—Hormone Replacement Therapy†: Progestin/Estrogen Reduces Incidence of CHD and Hip Fractures by Only 1/3 of the Designed Effect; Increases Incidence of Breast Cancer 80%; No Change in Deaths from Other Causes

	6 Years		z
	C (n = 6125)	I (n = 7000)	
Incidence^a			
CHD	3.26 (0.23)	3.04 (0.21)	0.72
Hip fractures	1.87 (0.17)	1.74 (0.16)	0.56
Breast Cancer	2.07 (0.18)	2.79 (0.20)	-2.69*
Endometrial Cancer	0.46 (0.09)	0.46 (0.08)	0.00
Mortality^b			
CHD	1.30 (0.14)	1.22 (0.13)	0.41
Hip fractures	0.47 (0.09)	0.44 (0.08)	0.25
Breast Cancer	0.52 (0.09)	0.70 (0.10)	-1.33
Endometrial Cancer	0.05 (0.03)	0.05 (0.03)	0.00
Other causes	5.37 (0.29)	5.37 (0.28)	0.00

†This scenario is based on the initial protocol. There was a major revision in the design of the HRT component in January 1995 that is not reflected here.

^aThe percentage of incident cases in each group (standard errors in brackets).

^bThe percentage of deaths in each group (standard errors in brackets).

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming.

also engendered substantial disagreement. In scenario 3 the dietary intervention appears not to change coronary heart disease incidence but to lead to significant reductions in the incidence of breast and colorectal cancer. Some members were reluctant to stop the trial early leaving this controversial result to throw doubt on the correct approach to coronary heart disease prevention. In scenario 6, some members did not wish to stop the trial early before resolving the question on colorectal cancer, arguing that colorectal cancer has more serious health consequences than osteoporosis.

Also shown in Table 10 are the results of several statistical stopping rules applied to each scenario. These methods have been described in the previous section. No statistical method accords exactly with the majority view of the members of the DSMC. As single monitoring measures, the primary outcomes (specific approach) corresponded more closely to the DSMC view than the global indices (global approach). However, a mixed approach requiring the primary outcome to be significant and the global index to be "supportive" corresponded even better with the DSMC view when the global index was an unweighted combination of disease endpoints.

The most notable divergence between the DSMC view and the statistical rules is in scenario 8, which portrays a significant excess of breast cancer in the estrogen/progestin group. Members of the DSMC were unanimous in the recommendation to stop, but all statistical methods except one led to a recommendation for continuation. The exception was the mixed approach that included anticipated adverse effects as a separate criterion for stopping. Generally, across all scenarios, this rule agreed most closely with the DSMC view.

Table 10 Opinions of DSMC Members and Results of Statistical Monitoring Methods for Each of the Eight Scenarios

	Scenario							
	1	2	3	4	5	6	7	8
DSMC Opinions								
Continue	9	8	3	6	12	3	3	0
Stop	2	2	7	5	0	7	5.5	12
Cannot decide	1	2	2	1	0	2	3.5	0
MAJORITY VOTE ^a	C	C	(S)	(C)	C	(S)	(S)	S
Statistical Methods								
(i) Primary Outcomes	C	C	S	S	C	S	C	C
(ii) Global Methods								
1. Total mortality	C	S	C	C	C	C	C	C
2. Unweighted combination	C	S	C	C	C	C	C	C
3. Weighted combination	C	S	C	C	C	C	C	C
4. Bayes weighted combination	C	S	C	C	C	C	C	C
(iii) Mixed Methods ^c								
1. Primary + global ^b signif.	C	C	C	C	C	C	C	C
2. Primary signif. + global ^b supportive	C	C	S	C	C	S	C	C
3. Primary signif. + global ^b supportive or Adverse effect signif.	C	C	S	S	C	S	S	S

^aC, continue; S, stop; () indicates committee is substantially divided.

^bUnweighted combination.

^cExplanation of Mixed Methods:

1. To stop we require primary and global outcomes to both be significant.
2. To stop we require primary to be significant and global to be supportive.
3. To stop we require either primary to be significant and global to be supportive or adverse effect to be significant.

^dOne individual who could not decide but was inclined to stopping assigned half a vote to each option!

CONCLUSION

In this paper we discuss in general terms the important and complex issues that underlie monitoring of a long-term prevention trial. We have outlined a general approach to formulating statistical guidelines for monitoring such trials and have described the use of this approach in the development of monitoring guidelines for the WHI clinical trial.

The results of the initial exercise with the DSMC for the WHI have led to some conclusions regarding the statistical methods we proposed. First, this process confirmed for us the deficiencies in relying on the primary outcome for monitoring. Even when the DSMC's opinion agreed with the monitoring rule, the reasoning behind their opinion often involved considerations addressed by the global measures and beyond the primary outcome. Second, the four global methods we considered led to identical recommendations in all scenarios. Either the scenarios themselves did not represent situations where incidence and mortality were strikingly different or else these approaches were all weighted quite consistently toward mortality. As such, the global methods used singly did not reflect the subtleties in balancing risks and benefits that either we or the DSMC identified, particularly for safety issues. We are more optimistic about the use of the mixed approach with separate monitoring of

adverse effects. This combination protects the test of the primary hypothesis while assuring that the broader safety and ethical issues of multiple diseases are systematically incorporated.

As further preparation for finalizing the guidelines, new exercises may be useful. First, since the DSMC members found it so helpful to discuss scenarios, we may devise further sets. Second, we may ask other groups of individuals to complete the scenario questionnaires, including investigators participating in the trial. Versions of the scenarios suitable for trial participants may also be prepared. Third, statistical methods that involve prediction of future results in the trial may be of particular use to cope with the problem of differential lag times to treatment effects. In the WHI clinical trial, there is particular concern over the possible risk of increased breast cancer incidence in women taking long-term hormone replacement therapy. Knowing whether to stop early in the face of beneficial effects on coronary heart disease, when the risks of increased breast cancer are still unknown, is particularly difficult, and the statistical methods described in this paper do not capture this aspect of the problem. We are presently investigating whether statistical prediction is a useful tool for handling this problem.

Other modifications to our statistical approach may deserve consideration. For example, as mentioned at the end of the "Mixed Approaches" section, instead of employing a single all-encompassing stopping rule, one may define a mixed approach rule for stopping because of benefit and another mixed approach rule for stopping because of an adverse effect. The two rules may then be applied in tandem. The mixed approach rule for stopping in response to adverse effects would allow one to balance adverse effects with any beneficial effects in deciding whether to stop the trial. Another possible modification is to define two levels of alert for the DSMC. A low-level alert would be signaled when evidence for an adverse effect emerges. The adverse effect may not be sufficiently serious or frequent to advocate stopping the trial, but the alert would initiate a debate on the need to inform the trial participants of the effect. A high-level alert would be signaled using a mixed approach stopping rule. The two-tiered level of alert would allow one to distinguish between the need to reinform participants and the need to stop the trial. All of the above considerations are shaping the guidelines for early termination that are being prepared for the WHI Clinical Trial. The full proposal will be presented to and debated by the DSMC at a future meeting.

Although we have developed our approach for prevention trials, it may also be useful for treatment trials. Trials of prophylaxis to prevent recurrence of a disease for which event rates are low may be particularly suited to this approach because low event rates increase the potential for effects on secondary diseases to influence the overall evaluation of the treatment. Even more generally, the approach could be used to attempt to bridge a gap between the formal stopping rules and the "real" problems of health care and patients that is sometimes felt to exist.

Some may argue that the value in our proposed approach lies more in the process than in the outcome. Preliminary discussion among members of the DSMC using scenarios is certainly valuable preparation for their task when the real data become available. Perhaps this is all that is needed. Perhaps hoping to formulate reasonable formal stopping guidelines on the basis of

statistical calculations is vain when the decision process is so complex. Why not simply use the usual type of stopping rules based on the primary outcome and leave the DSMC to deal with the remaining complexities as they arise? We would argue that, in fact, as the decision making process becomes more complex and the need for decisions becomes more pressing, committees tend to grasp for objective rules. The statistical guidelines actually assume considerable importance in these circumstances. Therefore, we advise investing serious effort in finding statistical guidelines that will represent, as far as possible, the best ethical and scientific monitoring of the trial. The methods described in this paper aim in that direction.

REFERENCES

1. Ellenberg S, Geller NL, Simon R, Yusuf S, eds. Practical issues in data monitoring of clinical trials. *Stat Med.* 1993;12:415-616.
2. Fleming TR, DeMets DL. Monitoring of clinical trials: Issues and recommendations. *Controlled Clin Trials.* 1993;14:183-197.
3. Canner PL. Monitoring of the data for evidence of adverse or beneficial effects. *Controlled Clin Trials.* 1983;4:467-484.
4. Smigel K. Breast Cancer Prevention Trial takes off [news]. *JNCI.* 1992;84:669-670.
5. Omenn GS, Goodman G, Thornquist M, Grizzle J, Rosenstock L, Barnhart S, Balmes J, Cherniack MG, Cullen MR, Glass A et al. The beta-carotene and retinol efficacy trial (CARET) for chemoprevention of lung cancer in high risk populations: smokers and asbestos-exposed workers. *Cancer Res.* 1994;54(Suppl):2038s-2043s.
6. *Women's Health Initiative Manuals. vol 1. Study Protocol and Policies.* WHI Clinical Coordinating Center, Fred Hutchinson Cancer Research Center, Version 1.0, Sept. 1, 1994, Seattle, WA, 1994.
7. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics.* 1979;35:549-556.
8. Geller NL, Pocock SJ. Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics.* 1987;43:213-224.
9. Spiegelhalter DJ, Freedman LS, Pannar MKB. Bayesian approaches to randomized trials. *J. R Stat Soc. Series A.* 1994;157:357-416.