CLINICAL TRIALS ARTICLE Clinical Trials 2007; 4: 207–217

Monitoring and reporting of the Women's Health Initiative randomized hormone therapy trials

Garnet L Anderson[a], Charles Kooperberg[a], Nancy Geller[b], Jacques E Rossouw[b], Mary Pettinger[a] and Ross L Prentice[a]

Background The Women's Health Initiative (WHI) randomized trial of estrogen plus progestin (E + P) was terminated early based on an assessment of harms exceeding benefits for disease prevention. The results contravened prevailing wisdom and a large body of literature regarding benefits of menopausal hormone therapy. The results and their interpretation have been the subject of considerable debate.

Purpose/methods To describe the process of developing a trial monitoring plan, the key interim and final data, and to explain the choice of statistical methods used in trial monitoring and reporting.

Results A formalized monitoring plan was developed using statistical methods that acknowledged protocol-defined design and analysis plans, input of monitoring board members especially regarding the role of various study outcomes, and multiple comparisons. Major early departures from design assumptions concerning treatment effects indicated a need for additional flexibility in safety monitoring. When the trials were stopped early, questions arose as to how closely the statistical methods in published reports should correspond to those defined by protocol or used in monitoring. Methods were selected to provide a simple and transparent summary of the primary results, with a cautious interpretation promoted by acknowledgement of multiple testing.

Conclusions Developing a formal trial monitoring plan with a view towards influencing clinical practice is useful for creating consensus among DSMB members regarding the evidence that would justify stopping a trial and the framework to be used to address statistical complexities. Departures from design assumptions typically occur. These reinforce the role of the DSMB in exercising their judgment, and the judicious adaptation of these statistical guidelines in monitoring and reporting trials. In communicating the results in such circumstances, priority should be given to presenting as fair, accurate and transparent a view of the data and findings as current methods and technology allow.

Clinical Trials 2007; **4**: 207–217. http://ctj.sagepub.com

Introduction

In July 2002, the National Heart Lung and Blood Institute (NHLBI) announced the early termination of the Women's Health Initiative (WHI) trial of estrogen plus progestin (E + P). This randomized, double-blind placebo-controlled trial of 16 608 postmenopausal women was stopped approximately three years early at the unanimous recommendation of the WHI Data and Safety Monitoring Board (DSMB), based on the assessment that health risks exceeded benefits for disease prevention in postmenopausal women over an average 5.2-year follow-up period [1]. The results contravened both prevailing wisdom and a large body of literature from observational studies, intermediate endpoint trials, and animal experiments [2].

There was an immediate and vocal response to this report [3,4], including discussion and criticism of the presentation and interpretation [5,6], which

[a]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [b]National Heart Lung Blood Institute, Bethesda, MD, USA

Author for correspondence: Garnet L Anderson, WHI Clinical Coordinating Center, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, M3-A410, PO Box 19024, Seattle, WA 98109, USA. E-mail: garnet@whi.org

 10.1177/1740774507079252

persisted for years [7–13]. Much of the debate can be attributed to the nature of the findings themselves, but some technical aspects continue to be the subject of discussion. Details of the statistical considerations that led to the stopping of the E + P trial and related choices that were made in presenting the results have not previously been published.

The principles that guided the reporting of the E + P trial were employed in reporting the parallel WHI trial of estrogen-alone (E-Alone) among 10 739 postmenopausal women less than two years later. The NHLBI terminated this trial in 2004 because of an increased risk of stroke and the fact that neither the hypothesized cardioprotective effect nor an adverse effect on breast cancer risk was likely to be demonstrated by continuing the intervention for the planned duration [14]. Because the results and stopping considerations for the E-Alone trial differed from those of the E + P trial, the application of these principles presented a different set of issues for consideration.

The present authors represent the unblinded statisticians at the WHI Clinical Coordinating Center (CCC) responsible for data analysis and reporting to the DSMB and unblinded members of the NHLBI Project Office responsible for trial oversight. In this article we briefly review statistical aspects of the design and provide further details of the trial monitoring plan. We summarize the data presented to the DSMB and their subsequent recommendations. We describe the implications of design and monitoring factors on the statistical presentations in the initial trial publications [1,14] and discuss considerations in these decisions and some lessons learned.

WHI design

The WHI is a large, multi-component public health research program, sponsored by the NHLBI, with input from other groups at the National Institutes of Health (NIH). The randomized clinical trial component of the WHI involved testing three intervention strategies for their effectiveness in chronic disease prevention in postmenopausal women: hormone therapy, a low-fat dietary modification, and calcium and vitamin D supplements. The hormone therapy component involved testing two different preparations in distinct subgroups of women: estrogen-alone in women with prior hysterectomy and estrogen plus progestin in women with an intact uterus. The primary objective of both trials was to determine whether hormone therapy would prevent coronary heart disease (CHD) and to provide an overall health benefit to postmenopausal women. Sample sizes for each trial were based on the CHD hypothesis. Breast cancer was the primary safety outcome. Several other outcomes were of interest, based on available literature at trial inception, but were viewed as secondary in trial motivation and design.

An underlying premise of the design was that there was a noteworthy lag-time to full effect of these hormones on the disease processes and hence, any early differences observed between randomized groups (eg, differences in breast cancer rates between active and placebo groups in the first year or two) would more likely arise from chance than from a true intervention effect. Under this assumption, power would be improved by down-weighting early differences. The analysis plans and sample size considerations were developed under this mindset using a weighted logrank statistic. The weighting function was defined by time since randomization, rising linearly from zero at randomization to one over a specified time interval (three years for cardiovascular disease (CVD) and fractures, ten years for cancer and mortality), and constant at one thereafter (Figure 1) [15].

WHI monitoring plan

The NHLBI Director appointed an independent DSMB to monitor and provide recommendations regarding the ethical conduct of the trial (membership provided in [1,14]). A formal trial monitoring plan was developed with input and final approval by the DSMB.

Monitoring plan development

The likelihood that multiple clinical outcomes would be affected by hormone therapy was a particular challenge for trial monitoring. To facilitate development of the monitoring plan, a rather unusual process was undertaken. Study statisticians from the CCC and NIH representatives developed

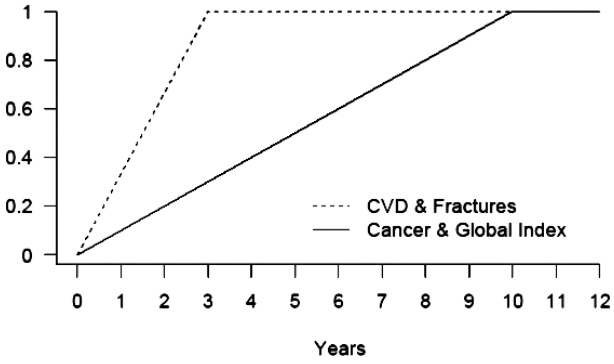

Figure 1 Weights defined by time since randomization as used in weighted logrank statistics for treatment arm comparisons of disease incidence rates.

scenarios projecting potential intervention comparisons for key outcomes at a point in time when approximately two thirds of the data had been acquired. The scenarios represented a variety of alternative hypotheses, including both higher and lower than anticipated benefits and risks. Members of the DSMB, and later study investigators, considered each scenario and indicated whether they would favor stopping or continuing the trial or whether they were unable to decide. The purpose of this exercise was to explore DSMB members' sensitivity to various intervention effects, to inform DSMB members on the mindset of their fellow members, and to use these to develop statistical guidelines in line with their judgement. Some aspects of this process have been published [16].

The DSMB responses to multiple scenarios suggested a heavy reliance on the primary outcome (CHD) and the primary safety outcome (breast cancer). Beneficial effects on several other outcomes were anticipated, but these were viewed by the DSMB as secondary for trial monitoring considerations, either because there was insufficient preliminary data to justify a trial to test that specific hypothesis (eg, hormone effects on stroke or colorectal cancer), or because trial results for that outcome were so highly anticipated that, viewed in isolation, they would not likely change clinical practice (eg, hormone effects on fractures). Consequently, the monitoring plan did not specify stopping boundaries for benefit (upper boundaries) for any secondary outcome.

Adverse effect (lower) boundaries were defined for all monitored outcomes. Death from other causes was included to capture serious but unforeseen adverse intervention effects.

Scenarios involving simultaneous evidence of both risks and benefits were specifically considered. In such a setting three potential actions were envisioned – stopping for benefit, stopping for harm, or continuing because there was insufficient clarity of overall benefits versus risks. A statistical summary of effects was sought to support this decision-making process. After consideration of several options, a so-called global index of risks and benefits was adopted, calculated for each woman as the time to the first event among the list of outcomes to be monitored.

Based on these considerations, the list of formally monitored outcomes for both trials was approved (Table 1) including the primary and key secondary outcomes identified in the protocol, plus colorectal cancer, another serious clinical event that had become of interest in relation to hormone therapy. Comparative data for a list of other clinical outcomes, as well as for a list of symptoms plausibly associated with hormone therapy, were provided biannually to the DSMB. Treatment arm comparisons of data from the WHI Memory Study [17], an ancillary study capturing information on probable dementia and mild cognitive impairment in a subset of trial participants, were also made available to inform early stopping discussions.

Table 1 Clinical outcomes subject to formal monitoring in the WHI Hormone Therapy Trial component

Outcomes	Estrogen + Progestin	Estrogen-Alone
Primary	Coronary heart disease	Coronary heart disease
Secondary	Hip fracture	Hip fracture
	Breast cancer*	Breast cancer*
	Colorectal cancer	Colorectal cancer
	Endometrial cancer	
	Stroke	Stroke
	Pulmonary embolism	Pulmonary embolism
	Death from other causes	Death from other causes
Global index	First occurrence of any of the above conditions	First occurrence of any of the above conditions

*Primary adverse outcome.

Conceptual framework for early stopping considerations

The monitoring plan indicated that early stopping considerations *for benefit* would be triggered only if the weighted logrank statistic for CHD crossed the CHD upper boundary and the global index was supportive of overall benefit, as assessed by comparison to a separate boundary. Benefits observed only on secondary outcomes were not expected to lead to early stopping discussions. Stopping for *adverse effects* would be considered if any disease-specific comparison crossed the associated lower boundary and the global index was suggestive of overall harm.

Although the trial motivation anticipated the same treatment effects for estrogen plus progestin and estrogen-alone, the design and monitoring of these distinct trials were done separately. Pooled results for the two trials were provided to the DSMB. The pooled analyses were expected to be useful if the data revealed a clear and early reduction in CHD incidence while suggesting an emerging adverse effect on breast cancer rates in both trials.

Statistical criteria specified in the monitoring plan

The monitoring plan was devised to limit the "experiment-wise" type I error rate using the traditional 0.05 level for benefit and the 0.10 level for adverse effects. The asymmetry in assessing benefits and risks was based on the importance of protecting study volunteers from research risks, a particular concern in prevention trials where participants are

ostensibly healthy. Stopping boundaries for all disease-specific comparisons were calculated according to the O'Brien-Fleming (OBF) procedure [18] for 15 interim semi-annual analyses, with the first analysis to occur in late 1997.

"Supportive" evidence of overall benefit was defined as the global index comparison exceeding a two-sided, 0.10-level OBF boundary. Evidence suggesting an overall adverse effect used a less stringent criterion for the lower boundary, defined by a standardized, normally-distributed test statistic (*Z*-score) of less than −1.0. Viewed in isolation, the boundaries for the global index were not conservative. However, the global index was not intended to be used as a stand-alone monitoring tool; rather, the monitoring plan called for considering the global index only when an individual disease comparison crossed its corresponding boundary. As a second level criterion, the global index introduced additional conservatism into the monitoring boundaries beyond that of the OBF procedures.

No adjustment for multiple outcomes was incorporated in the monitoring boundaries for CHD benefit or breast cancer adverse effect. Stopping boundaries for other adverse effects used a Bonferroni correction based on six outcomes for E-alone (CHD, stroke, pulmonary embolism, colorectal cancer, endometrial cancer, hip fractures, and death from other causes) and seven outcomes for E + P (all above plus endometrial cancer).

The substantial difference in anticipated lag-time to full intervention effect between the primary outcome of CHD (three years) and the primary safety outcome of breast cancer (10 years) was a specific concern. The question as to the appropriate action to take if the boundary for CHD benefit were crossed early but the breast cancer comparison strongly suggested harm was difficult. No explicit statistical method was devised to address this issue, other than the conservatism built into the stopping guidelines through the global index.

In summary, a discussion of stopping for benefit would be triggered only if both the upper 0.05-level boundary for CHD and the upper 0.10-level boundary for the global index were crossed. Stopping for an overall adverse effect would be considered if any of the disease-specific lower boundaries were crossed (0.10-level for breast cancer or the Bonferroni corrected 0.10-level for other listed outcomes) and the global index logrank statistic was less than −1.0. The stopping boundaries and logrank statistics for key outcomes of each trial were presented on the standardized normal scale, as illustrated in Figures 2 and 3 for the E + P and E-Alone trials, respectively. The final plan was formally approved by the DSMB in early 1998 after one official interim analysis had been conducted, but had been in essentially final form throughout the preceding year.

Monitoring and early stopping of the hormone therapy trials

Issues arising during monitoring

In August 1998, the Heart Estrogen/Progestin Replacement Study (HERS), a secondary prevention trial of the same E + P preparation in women with established coronary disease, presented its principal results and reported no overall effect on CHD event rates after 4.1 years of follow-up, but a statistically significant elevation in coronary disease in the first year of exposure as well as an overall elevation in venous thromboembolism (VT) [19]. Interim analyses of the WHI hormone therapy trial data at that time suggested similar early adverse effects on CHD, stroke (Figures 2 and 3) and VT (data not shown). The weighted logrank statistics, which discounted early differences, lacked sensitivity to these early effects. Because this was a safety concern, the DSMB requested that the monitoring reports be augmented with unweighted logrank statistics as an aid to early adverse effect monitoring. All subsequent reports contained both weighted and unweighted logrank statistics.

By early 2000, interim treatment arm comparisons for CHD, stroke, and pulmonary embolism strongly suggested early risk elevations in both trials with nominally statistically significant comparisons in the E + P trial for CHD and stroke (unweighted $Z < -1.96$, Figure 2) and for pulmonary embolism (data not shown). No stopping boundaries had been crossed though, other than the lower boundary for the global index. The DSMB recommended that participants in both trials be informed of these early, unanticipated risks. There were differences in the magnitude of effects between the two trials, but since the estimated hazard ratios (HR) for cardiovascular diseases were elevated in both, all hormone trial participants were informed of this early elevation in heart disease, strokes and blood clots that may diminish with time. In 2001, as evidence of these adverse effects persisted the DSMB recommended that participants again be told of the potential adverse cardiovascular effects. WHI investigators informed participants of these risks without quantitative information or unblinding and with a statement that the DSMB recommended the trials continue (participant materials available [20]).

Stopping of the E + P trial

In early April of 2002, analyses of data available through 28 February 2002 revealed that the breast cancer statistic in the E + P trial (weighted logrank $Z = -3.19$) had crossed the adverse effect boundary ($Z = -2.32$) and the global index statistic (weighted

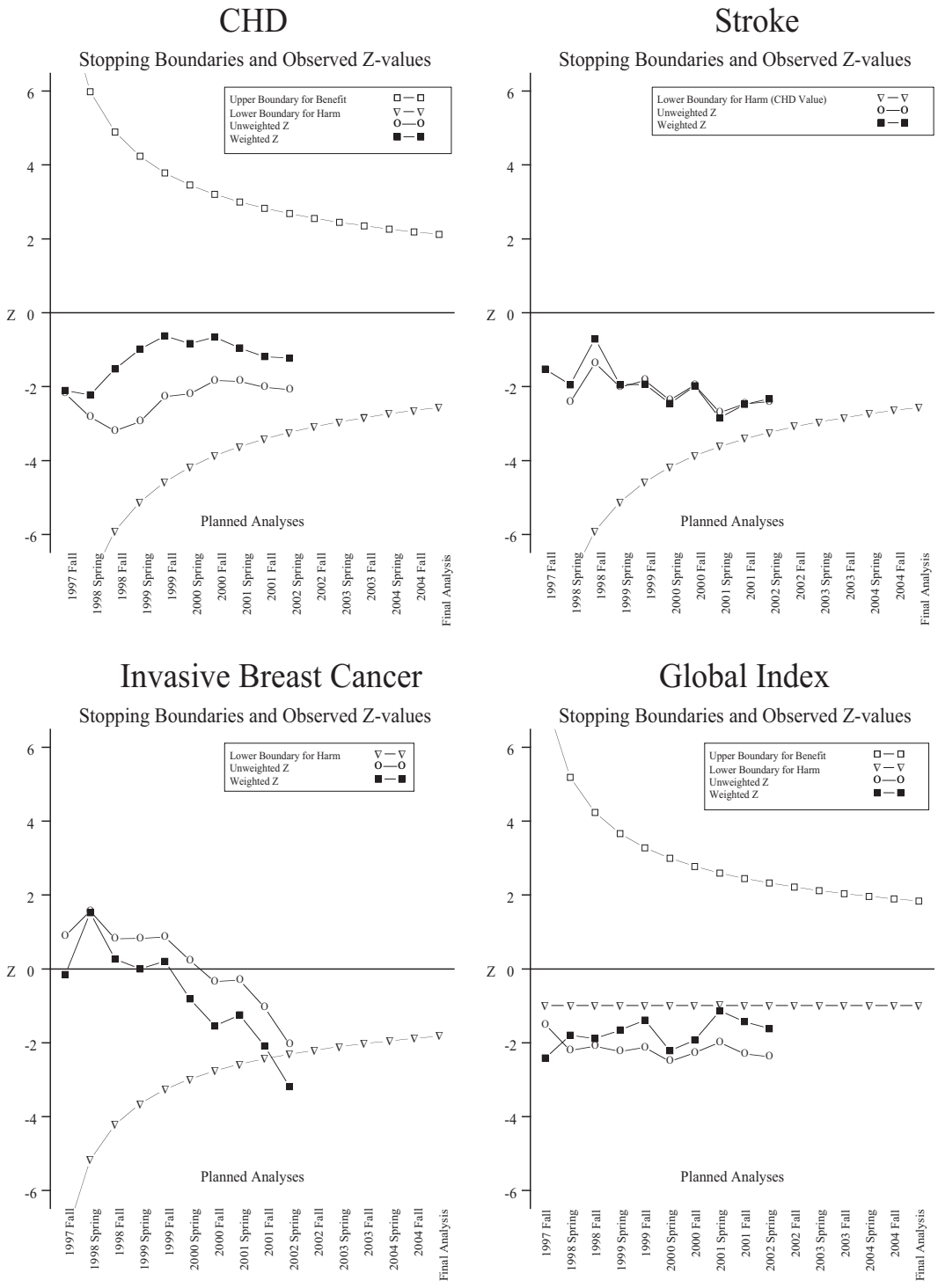

Figure 2 Monitoring boundaries and interim logrank statistics for selected outcomes in the WHI E + P trial. According to the monitoring plan, a weighted logrank (Z) statistic above (below) the corresponding monitoring boundary provides evidence for stopping for benefit (adverse effect). Unweighted logrank statistics were to be considered when examining potential early adverse effects. Stopping considerations were to be based on evidence of an effect for a specific disease, supported by the global index exceeding the its corresponding monitoring boundary in the same direction. The breast cancer weighted logrank statistic exceeded the adverse effect boundary in Spring 2002, and a global index comparison which was considered supportive of an overall adverse effect throughout the trial. CHD and stroke comparisons showed nominally significant adverse effects (corresponding unweighted $Z<-1.96$) but did not cross monitoring boundaries.

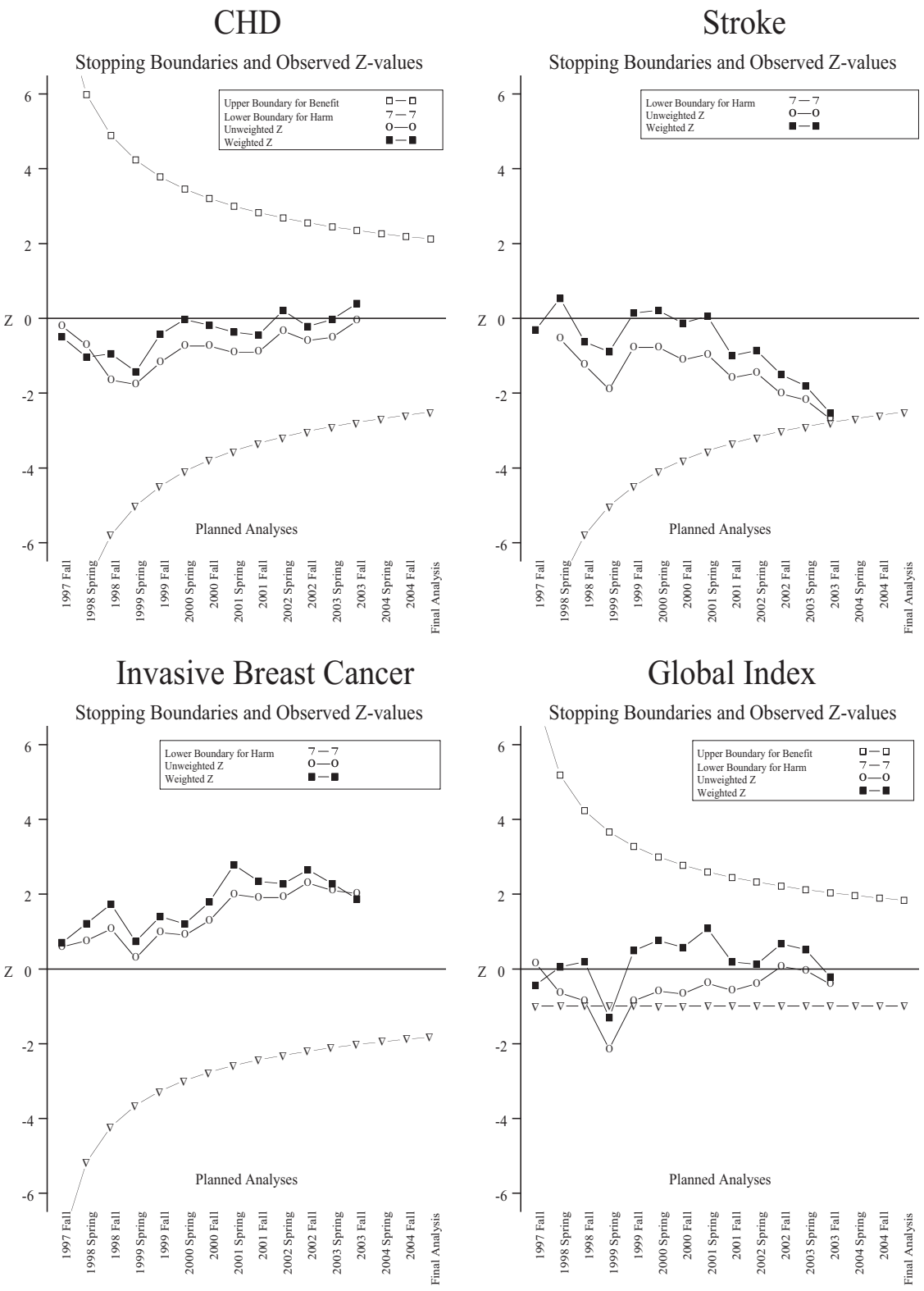

Figure 3 Monitoring boundaries and interim logrank statistics for selected outcomes in the WHI E-Alone trial. In the fall of 2003, the stroke comparison was approaching the adverse effect boundary, the breast cancer comparison had reached nominal statistical significance for benefit (unweighted $Z > 1.96$) and the CHD and global index comparisons were neutral.

Table 2 Tenth interim analysis of the WHI E + P Trial[a]. (Data as of: 28 February 2002.)

	E + P		Placebo				
Number randomized	8506		8102				
Mean follow-up time (months)	60.2		59.2				
Outcomes	*n*	%	*n*	%	Weighted *Z*	Unweighted *Z*	Hazard ratio[b]
Coronary heart disease (CHD)	160	1.88	119	1.47	−1.23	−2.09	1.29
Hip fractures	42	0.49	61	0.75	1.89	2.20	0.64
Invasive breast cancer	160	1.88	116	1.43	−3.19	−2.05	1.28
Endometrial cancer	22	0.26	24	0.30	1.00	0.50	0.86
Colorectal cancer	41	0.48	63	0.78	2.73	2.51	0.61
Stroke	120	1.41	80	0.99	−2.33	−2.41	1.41
Pulmonary embolism	66	0.78	31	0.38	−2.17	−3.23	1.99
Death from other causes[c]	152	1.79	153	1.89	0.81	0.76	0.92
Global index[d]	713	8.38	592	7.31	−1.62	−2.38	1.14
Total deaths	214	2.52	203	2.51	0.58	0.24	0.98

[a]Monitoring boundaries for the 10th interim analyses were:

CHD benefit	$Z = 2.68$	Upper boundary
Global index benefit	$Z = 2.32$	Upper boundary
Breast cancer adverse effect	$Z = -2.32$	Lower boundary
Other adverse effects	$Z = -3.25$	Lower boundary
Global index for harm	$Z = -1.00$	Lower boundary, in conjunction with crossing a stopping boundary for breast cancer or other adverse effects

[b]From an unweighted proportional hazards model stratified by age, prevalent condition, and DM randomization arm.
[c]All deaths except those from breast, colorectal, or endometrial cancer, CHD, stroke, pulmonary embolism or hip fracture. Includes deaths that are not yet adjudicated.
[d]Participants having one or more of the above listed outcomes.

$Z = -1.62$, unweighted $Z = -2.38$) was supportive of overall harm (Table 2). CCC statisticians informed the unblinded NHLBI Project Officer and together developed a contingency plan for early stopping. Because the scheduled DSMB meeting was approximately eight weeks later, the first step was to conduct an unscheduled analysis of data accrued through 30 April 2002. The purpose of this analysis was to confirm that these statistics remained beyond the boundaries, providing additional assurance against a chance finding. The updated analyses, available by mid-May, produced a breast cancer comparison that was slightly attenuated ($Z = -2.92$) but still exceeded the stopping boundary. Supportive evidence of overall harm also persisted (global index unweighted $Z = -1.74$). Analyses from both time points were presented to the DSMB at the regular meeting on 31 May. Based on these data, which met the established adverse effect stopping criteria, the DSMB recommended that the E + P trial be stopped. The NHLBI Director attended the meeting and accepted the recommendation. Contingency plans for stopping the trial were launched immediately. An article describing the data available through 30 April was submitted on 5 June and published in mid-July [1]. Simultaneously all 27 000 participants in the two hormone trials were notified by mail of these results.

Stopping of the E-Alone trial

Despite the early stopping of the E + P trial in 2002, the DSMB recommended that the E-Alone trial continue as planned. No stopping boundaries had been crossed and the overall risk/benefit profile was approximately neutral. In November 2003, at the regularly scheduled meeting of the DSMB, and in subsequent conference calls, the committee could not reach a consensus recommendation for continuing the trial. No stopping boundaries had yet been crossed, but the stroke hazard ratio was similar to that observed in the E + P trial and the comparison was near the adverse effect stopping boundary. The global index was balanced, owing to hazard ratios less than one for hip fracture and unexpectedly for breast cancer (Table 3 and Figure 3). Because the DSMB was undecided, the NHLBI constituted a separate committee to review the results and in February 2004 stopped the E-Alone trial based on the elevated risk of stroke and lack of CHD benefit [21].

Statistical issues in reporting results from terminated trials

In reporting these results, the question arose as to how closely the data presentations to the medical

Table 3 Thirteenth interim analysis of the WHI E-Alone Trial.[a] (Data as of 31 August 2003.)

	E-Alone		Placebo				
Number	5310		5429				
Mean f	76.3		76.6				
Outcomes	*n*	%	*n*	%	Weighted Z	Unweighted Z	Hazard ratio[b]
Coronary heart disease	166	3.13	169	3.11	0.39	−0.07	1.01
Hip fractures	29	0.55	57	1.05	2.74	2.93	0.52
Invasive breast cancer	87	1.64	118	2.17	1.87	2.02	0.75
Colorectal cancer	57	1.07	55	1.01	−0.80	−0.35	1.07
Stroke	144	2.71	106	1.95	−2.54	−2.69	1.41
Pulmonary embolism	40	0.75	29	0.53	−1.23	−1.43	1.42
Death from other causes[c]	170	3.20	161	2.97	−0.67	−0.82	1.09
Global index[d]	623	11.73	623	11.48	−0.22	−0.41	1.02
Total deaths	258	4.86	248	4.57	−0.11	−0.80	1.07

[a]Monitoring boundaries for the 13th interim analyses were:

CHD benefit	$Z = 2.35$	Upper boundary
Global index benefit	$Z = 2.03$	Upper boundary
Breast cancer adverse effect	$Z = -2.03$	Lower boundary
Other adverse effects	$Z = -2.79$	Lower boundary
Global index for harm	$Z = -1.00$	Lower boundary, in conjunction with crossing a stopping boundary for breast cancer or other adverse effects

[b]From an unweighted proportional hazards model stratified by age, prevalent condition and DM randomization arm.
[c]All deaths except those from breast or colorectal cancer, CHD, stroke, pulmonary embolism or hip fracture. Includes deaths that are not yet adjudicated.
[d]Participants having one or more of the above listed outcomes.

community and the public should correspond to those defined by protocol or used in trial monitoring. Three technical issues were raised in this regard: Should the results be reported with weighted or unweighted test statistics, and should the same approach be used for all outcomes? Should the multiple interim analyses and multiple outcomes be acknowledged in confidence interval formulation and if so, how? And, should the asymmetry in monitoring benefits and adverse effects carry over into reporting? The decisions described below were made for the E + P trial, and the general approach was subsequently applied to the E-Alone trial.

Unweighted and weighted test statistics

The protocol specified weighted logrank statistics for each clinical outcome but an unweighted Cox regression analysis was used in publishing E + P results [1], motivated by the importance of providing a transparent and easily interpretable (hazard ratio) estimate of treatment effects. The unweighted analysis was particularly preferred for reporting of the CHD result because it did not rely on the assumption of lag time to full prevention effect, an assumption that was not supported by the data. For simplicity, the same methods were used for reporting all other outcomes, even though this provided a more conservative breast cancer statistic, compared to the corresponding weighted test. The weighted logrank statistic for breast cancer, the catalyst for stopping the trial, was highly statistically significant ($P = 0.001$), but the lower limit of the 95% confidence interval from the unweighted analyses was one (HR: 1.26; 95% CI: 1.00–1.59) [1]. The weighted test significance level was also presented but only in the description of the trial termination. The strength of these findings was questioned by some [22,23] until subsequent analyses of breast cancer rates relying on the protocol-defined statistics were published [24].

Acknowledging multiple comparisons

The initial E + P trial report provided both nominal and multiple comparisons-adjusted confidence intervals (CIs), the latter included to discourage over-interpretation and to put the results into the context in which the stopping decision was made. The adjustments to the width of the CIs were made by an appropriate transformation of the monitoring boundaries. The use of the OBF procedures and Bonferroni adjustments to correct confidence intervals for multiplicities provides a very conservative view of these data, particularly in reporting secondary outcomes where both adjustments were applied. While this conservatism seems appropriate for trial monitoring where early stopping decisions of trials involving such an enormous investment need to be considered with due caution, it likely

under-represents the true statistical significance of the results. Any other adjustment for multiple comparisons developed when results were in hand, however, would be unduly *ad hoc*.

These two sets of CIs provide somewhat different inference for individual disease effects, but can be viewed as bracketing the likely true confidence region limits. For example, in interpreting the effects of E + P on CHD, where the reported E + P hazard ratio was 1.29, the nominal CI of 1.02–1.63 can be interpreted as providing evidence of a modest increase in CHD over 5.2 years. The adjusted CI of 0.85–1.97 does not allow one to conclude there is an increase in CHD rates, but does exclude the level of benefit the trial was designed to detect (a 22% reduction). In this case, the adjusted CI is likely too conservative. Specifically, the OBF adjustment assumes that early stopping is based on the CHD statistic crossing its prespecified boundary. Here, however, the CHD data played only a minor role in the early stopping decision so that the nominal confidence interval at that time can be expected to be fairly accurate.

Ignoring the asymmetry in risk and benefit evaluation

The monitoring plan required stronger statistical evidence of benefit (two-sided P-value < 0.05) than for adverse effects (two-sided P-value < 0.10). To incorporate this asymmetry into confidence intervals would have been both awkward and inconsistent with reporting standards, so the wider 95% CIs were used uniformly for all adverse effects.

Additional issues in reporting

The 2002 E + P publication included both relative and absolute risk estimates of treatment effects for all outcomes. Use of relative risk estimates was consistent with the planned analyses and motivated by their well-established statistical properties, including the ease of accounting for stratification factors and the somewhat better control for follow-up time, yielding well-behaved tests and estimates while avoiding strong modeling assumptions. Absolute risk calculations, based on simple differences in annualized incidence rates, were included to facilitate clinical interpretation.

Application of these principles to the E-Alone trial

Because the termination of the E-Alone trial was not triggered by fulfilling the stopping criteria of the monitoring plan, the extent to which the adjustments for multiple analyses should be acknowledged is debatable. The E-Alone trial termination was not independent of the multiple interim analyses or multiple disease event rate comparisons, however, so a parallel approach was taken in presenting these results. The reliance on unweighted statistics for this trial was less of a concern. With the longer-term follow-up (average of 6.8 years for E-Alone), which was approaching the planned termination, the differences between weighted and unweighted analyses were smaller. Further, the wide interest in comparing these results with the E + P trial findings provided strong impetus to use the same analytic techniques.

Comment

The WHI trial monitoring plan provided the foundation of analyses and reporting to the WHI DSMB and gave guidance to their assessment of trial evidence. The plan was informed by *a priori* expectations of benefits and risks of hormone therapy and developed around key design assumptions, the protocol-defined analysis plan, and standard statistical methods for multiple outcomes and interim analyses. It was tailored to support the clinical and ethical judgment of the DSMB by an iterative process of assessing the DSMB's responses to hypothetical scenarios of interim results. This development process was instrumental in bringing forward various perspectives of board members in their assessment of the clinical environment in which these data would be viewed. The statistical framework for the stopping boundaries was derived to present the data in alignment with their judgement.

The concept of a global index of risks and benefits was one particularly useful aspect of the plan that arose from these discussions. The primary role of the global index in this plan was to promote caution in early termination if there was emerging evidence in both directions. Possible trade-offs in treatment effects exist in many settings (eg, AIDS, cancer therapy) but there is heightened sensitivity to harm in chemoprevention trials conducted among ostensibly healthy individuals [25,26].

Evaluating individual treatment effects across a range of diseases to assess overall impact is not straightforward if they vary in incidence, morbidity and mortality; time-dependent treatment effects on these disease processes add further complication. In such circumstances, additional statistical summaries can be helpful. In these hormone trials, total mortality was considered a valid summary but too insensitive to effects on the chronic diseases being tested. A summation of the disease specific incidence rates, including death from other causes, seemed preferable. To acknowledge disease burden disparity across an initial list of outcomes (eg, vertebral fractures and deep vein thrombosis versus

stroke and colorectal cancer), the global index was first proposed as a weighted linear combination of disease incidence with weights defined by a measure of disease burden. The difficulty in determining appropriate weights and the statistical issues associated with multiple events per individual led us to instead define the global index as an unweighted summary using only the relevant clinical outcomes considered to be serious, life-threatening conditions (including death from other causes) using a time to first event data summary. Perhaps the most important aspect of the WHI global index was that it was specified in advance, preventing it from being skewed by selecting outcomes to be included based on observed results.

The monitoring plan provided the basic framework and guidance for monitoring, but the DSMB was not limited to its criteria in their considerations. Some of the emerging data in these and in other hormone therapy trials were inconsistent with design assumptions regarding cardiovascular effects, both in direction and timeline. Changing the monitoring plan based on emerging data can inflate the experimental error rates, however, so such modifications must be approached with caution. Here, the ethical requirements of safety monitoring and the information from other studies were sufficient to warrant a modest change, accomplished in this case by augmenting the protocol-defined weighted analyses with unweighted analyses. Other proposed modifications, including redefining outcomes to be monitored, were not adopted because of their likely effect on inflating the type I error since the justification arose primarily from observations within the trial.

In determining how to present the results when the E + P trial was stopped, the statistical approaches were chosen to balance several factors: consistency with the original design; analysis and monitoring plans; transparency and accessibility of results; and accuracy in portraying the data underlying the stopping decision. The departures from expectations and the complexities of the trial design and monitoring made it difficult to put these results into the appropriate context while avoiding extensive statistical considerations.

The initial publications of both trials [1,14] relied on traditional analytic approaches (i.e., unweighted analyses) for transparency and simplicity, even though the protocol-defined analysis for E + P and breast cancer would have suggested higher statistical significance. Use of the more conservative approach was expected to discourage over-interpretation, but it also created room for doubts as to whether the breast cancer finding was real. Inclusion of the protocol-defined analysis in the data displays might have prevented this. We presented both nominal and multiple comparison-adjusted CIs. While the latter provided another source of caution, interpretation was clearly hampered by including two sets of CIs. This remains an area that would benefit from additional statistical methods development [27] to more accurately describe the statistical significance of results in the presence of such multiplicities.

Our observations of WHI trial monitoring and reporting have reinforced the need for as much formality as is practical in the statistical analysis and monitoring plan. In particular, the specification of outcomes, statistical methods and early stopping procedures in advance promote clarity in the process and may help to avoid possible outside distortion of findings by those having strong prior beliefs or special interests. This formality in the monitoring guidelines does not diminish the importance of the clinical and ethical judgement of DSMB members nor bind them to the methods specified. Rather, it provides a useful basis from which the need for modifications can be assessed. In fact, the process of developing a formal monitoring plan in the multidisciplinary atmosphere of a DSMB is helpful in establishing from the outset an understanding of the evidence required to change clinical practice.

Communication of risk estimates remains a difficult area. Both relative and absolute risk estimates were presented, each with their own strengths and weaknesses. Whether these risks appear small or not, may be influenced by the metric (eg, a 29% increase in CHD in relative terms, or seven additional cases per 10 000 person-years of exposure). Critics have focused on the difference in perception of risk for the adverse effects [22, 23] without a parallel comment on the assessment of benefits (eg, a 34% reduction in hip fracture rates represented five fewer cases per 10 000 person years). The experience of WHI suggests that both types of estimates are useful – relative risk for inferences regarding within populations comparisons and absolute risk estimates for individual level risk assessment – and the use of both contributes to a balanced perspective.

A comprehensive assessment of the risks and benefits of a chronic disease prevention intervention remains a medical, social, and statistical challenge as more recent experiences with the Cox-2 inhibitor trials reveal [28]. Because the results are likely to be translated to a broad population, it is important that the assessment of effects be sufficiently comprehensive in scope. It is important that there be adequate power to detect reasonable levels of effects on the more common diseases in the population likely to be exposed. It is also important that the assessment of effects take into account the potential for duration-dependent effects. The WHI was designed to be such a program. The statistical challenges described here in monitoring and

presenting these unanticipated results are not unique to hormone therapy. Rather, these likely reflect the difficulty in changing clinical practice when it has embraced treatments or interventions in advance of a highly reliable assessment [29]. In reporting these results, our experience suggests that deviations from the usual approaches, even when intended to inject a measure of caution in the interpretation, may lead to some confusion or criticism. It seems important, however, not to let this possibility deter one from presenting as fair and accurate a view of the data and findings as current methodology and technology allow.

Acknowledgements

The authors would like to thank Drs Janet Wittes and Gerardo Heiss for their reviews of the manuscript and Ms Jenny Schoenberg for assistance with developing the data displays. This work was supported by a contract from the National Heart Lung and Blood Institute. Dr Prentice's contribution was partially supported by NIH grant CA53996.

References

1. **Writing Group for the Women's Health Initiative.** Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results of the Women's Health Initiative randomized controlled trial. *JAMA* 2002; **288**: 321–33.
2. **Derry PS.** Hormones, menopause, and heart disease: making sense of the Women's Health Initiative. *Women's Health Issues* 2004; **14**(6): 212–19.
3. The truth about hormones [cover story]. *Time*, 22 July 2002.
4. The end of the age of estrogen [cover story]. *Newsweek* 22 July 2002.
5. **Strickler RC**. Women's Health Initiative results: a glass more empty than full. *Fertility and Sterility* 2003; **80**: 488–90.
6. **Grimes DA, Lobo DA.** Perspectives on the Women's Health Initiative trial of hormone replacement therapy. *Obstetrics and Gynecology* 2002; **100**: 1344–53.
7. **Gambrell RD.** The Women's Health Initiative Reports in perspective: facts or fallacies? *Climacteric* 2004; **7**: 225–28.
8. **Stampfer M.** Commentary: Hormones and heart disease: do trials and observational studies address different questions? *Int J Epidemiol* 2004; **33**: 454–55.
9. **Vandenbroucke JP.** Commentary: The HRT story: vindication of old epidemiological theory. *Int J Epidemiol* 2004; **33**: 456–57.
10. **Barrett-Connor E.** Commentary: Observation versus intervention—what's different? *Int J Epidemiol* 2004; **33**: 457–59.
11. **Kuller LH.** Commentary: Hazards of studying women: the oestrogen oestrogen/progesterone dilemma. *Int J Epidemiol* 2004; **33**: 459–60.
12. **Petitti D.** Commentary: Hormone replacement therapy and coronary heart disease: four lessons. *Int J Epidemiol* 2004; **33**: 461–63.
13. **Lawlor DA, Smigh GD, Ebrahim S.** Commentary: The hormone replacement—coronary heart disease conundrum: is this the death of observational epidemiology? *Int J Epidemiol* 2004; **33**: 464–67.
14. **Women's Health Initiative Steering Committee.** Effects of conjugated equine estrogens on postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. *JAMA* 2004; **291**: 1701–12.
15. **The WHI Study Group.** Design of the WHI clinical trial and observational study. *Control Clin Trials* 1998; **19**: 61–109.
16. **Freedman L, Anderson GL, Kipnis V** *et al*. Approaches to monitoring the results of long-term disease prevention trials. *Control Clin Trials* 1996; **17**(6): 509–25.
17. **Shumaker SA, Reboussin BA, Espeland MA** *et al*. WHIMS: A trial of the effect of estrogen therapy in preventing and slowing the progression of dementia. *Control Clin Trials* 1998; **19**: 604–21.
18. **O'Brien PC, Fleming RT.** A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**: 549–56.
19. **Hulley S, Grady D, Bush T** *et al*. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA* 1998; **280**: 605–13.
20. Participant updates. Available at http://www.whiscience.org/about/about_ht.php. Last accessed 18 December 2006.
21. **Alving B.** NIH press statement, 2 March 2004. Available at http://www.nhlbi.nih.gov/new/press/04-03-02.htm. Last accessed 18 December 2006.
22. **Goodman N, Goldzieher J, Ayala C.** Critique of the report from the writing group of the WHI. *Menopausal Medicine* 2003; **10**: 1–4.
23. **Burger H.** Meeting report: Hormone replacement therapy in the post-Women's Health Initiative era. *Climateric* 2003; **6**(Suppl): 13–16.
24. **Chlebowski RT, Hendrix SL, Langer RD** *et al*. Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women: the Women's Health Initiative randomized trial. *JAMA* **203**: 32430–53.
25. **Anderson GL, Prentice RL.** Individually randomized intervention trials for disease prevention and control. *Statistical Methods in Medical Research* 1999; **8**: 287–309.
26. **Albanes D, Heinonen OP, Taylor PR** Alpha-Tocopherol and beta-carotene supplements and lung cancer incidence in the alpha tocopherol, beta-carotene cancer prevention study: effects of base-line characteristics and study compliance. *J Natl Canc Inst* 1996; **88**: 1560–70.
27. **Prentice RL, Pettinger M, Anderson GL.** Statistical issues arising in the Women's Health Initiative. *Biometrics* 2005; **61**: 899–941.
28. FDA Public Health Advisory: Non-Steroidal Anti-Inflammatory Drug Products (NSAIDS). 23 December 2004. Available at http://www.fda.gov/cder/drug/advisory/nsaids.htm.
29. **Hemminki E.** Opposition to unpopular research results: Finnish professional reactions to the WHI findings. *Health Policy* 2004; **69**: 283–91.