Biometrics 61, 899–941
December 2005
DOI: 10.1111/j.1541-0420.2005.00454.x

Statistical Issues Arising in the Women's Health Initiative

Ross L. Prentice,* Mary Pettinger, and Garnet L. Anderson*****

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center,
P.O. Box 19024, Seattle, Washington 98109-1024, U.S.A.
**email:* rprentic@whi.org
***email:* mpetting@whi.org
****email:* garnet@whi.org

Summary. A brief overview of the design of the Women's Health Initiative (WHI) clinical trial and observational study is provided along with a summary of results from the postmenopausal hormone therapy clinical trial components. Since its inception in 1992, the WHI has encountered a number of statistical issues where further methodology developments are needed. These include measurement error modeling and analysis procedures for dietary and physical activity assessment; clinical trial monitoring methods when treatments may affect multiple clinical outcomes, either beneficially or adversely; study design and analysis procedures for high-dimensional genomic and proteomic data; and failure time data analysis procedures when treatment group hazard ratios are time dependent. This final topic seems important in resolving the discrepancy between WHI clinical trial and observational study results on postmenopausal hormone therapy and cardiovascular disease.

Key words: Chronic disease prevention; Clinical trial monitoring; Genome-wide scan; Hazard ratio; Measurement error; Nutritional epidemiology; Observational study; Randomized controlled trial; Women's health.

1. Introduction

The Women's Health Initiative (WHI) is perhaps the most ambitious population research investigation ever undertaken. The centerpiece of the WHI program is a randomized, controlled clinical trial (CT) to evaluate the health benefits and risks of four distinct interventions (dietary modification, two postmenopausal hormone therapy [HT] interventions, and calcium/vitamin D supplementation) among 68,132 post-menopausal women in the age range 50–79 at randomization. Participating women were identified from the general population living in proximity to any of the 40 participating clinical centers throughout the United States. The WHI program also includes an observational study (OS) that comprised 93,676 postmenopausal women recruited from the same population base as the CT. Enrollment into WHI began in 1993 and concluded in 1998. Intervention activities in the estrogen plus progestin HT component of the CT ended early on July 8, 2002 when evidence had accumulated that the risks exceed the benefits. Intervention activities in the estrogen-alone component of the CT also ended early, on February 29, 2004. Intervention activities in the other two CT components ended on March 31, 2005. Nonintervention follow-up on participating women is planned through 2010, giving an average follow-up duration of about 13 years in the CT and 12 years in the OS.

The CT used a "partial factorial" design. Participating women met eligibility for, and agreed to be randomized to, either the dietary modification (DM) or one of the HT components, or both the DM and HT. The DM component randomly assigned 48,835 eligible women to either a sustained low-fat eating pattern (40%) or self-selected dietary behavior (60%), with breast cancer and colorectal cancer as designated primary outcomes and coronary heart disease (CHD) as a secondary outcome. The nutrition goals for women assigned to the DM intervention group were to reduce total dietary fat to 20%, and saturated fat to 7%, of corresponding daily calories and, secondarily, to increase daily servings of vegetables and fruits to at least five and of grain products to at least six, and to maintain these changes throughout the trial intervention period. The randomization of 40%, rather than 50%, of participating women to the DM intervention group was intended to reduce trial costs, while testing trial hypotheses with specified power.

The postmenopausal HT clinical trial components comprised two parallel randomized, double-blind, placebo-controlled trials among 27,347 women, with CHD as the primary outcome, with hip and other fractures as secondary outcomes, and with breast cancer as a primary adverse outcome. Of these, 10,739 women (39.3% of total) had a hysterectomy prior to randomization, in which case there was a randomized allocation between conjugated equine estrogen (E-alone) 0.625 mg/day or placebo. The remaining 16,608 (60.7%) of women, each having a uterus at baseline, were randomized (aside from an early assignment of 331 of these women to E-alone) to the same preparation of estrogen plus 2.5 mg/day of medroxyprogesterone (E+P) or placebo. A total of 8050 women were randomized to both the DM and HT clinical trial components.

At their 1-year anniversary from DM and/or HT trial enrollment, all CT women were further screened for possible randomization in the calcium and vitamin D (CaD) component, a randomized, double-blind, placebo-controlled trial of 1000 mg elemental calcium plus 400 international units of vitamin D_3 daily, versus placebo. Hip fracture is the designated primary outcome for the CaD component, with other fractures and colorectal cancer as secondary outcomes. A total of 36,282 (53.3% of CT enrollees) were randomized to the CaD component.

The total CT sample size of 68,132 is only 60.6% of the sum of the individual sample sizes for the four CT components, providing a cost and logistics justification for the use of a partial factorial design with overlapping components.

Postmenopausal women of ages 50–79 years who were screened for the CT but proved to be ineligible or unwilling to be randomized were offered the opportunity to enroll in the OS. The OS is intended to provide additional knowledge about risk factors for a range of diseases, including cancer, cardiovascular disease, and fractures. It has an emphasis on biological markers of disease risk, and on risk factor changes as modifiers of risk.

There was an emphasis on the recruitment of women of racial/ethnic minority groups throughout the WHI. Overall, 18.5% of CT women and 16.7% of OS women identified themselves as other than white. These fractions allow meaningful study of disease risk factors within certain minority groups in the OS. Also, key CT subsamples are weighted heavily in favor of the inclusion of minority women in order to strengthen the study of intervention effects on specific intermediate outcomes (e.g., changes in blood lipids or micronutrients) within minority groups.

To ensure adequate power for principle outcome comparisons, age distribution goals were specified for the CT as follows: 10%, ages 50–54 years; 20%, ages 55–59 years; 45%, ages 60–69 years; and 25%, ages 70–79 years. While there was substantial interest in assessing the benefits and risks of each CT intervention over the entire 50–79 year age range, there was also interest in having a sufficient representation of younger (50–54 years) postmenopausal women for meaningful age group-specific intermediate outcome (biomarker) studies, and of older (70–79 years) women for studies of treatment effects on quality of life measures, including aspects of physical and cognitive functioning. Differing shapes for age incidence rate functions within the 50–79 age range across the clinical outcomes that were hypothesized to be affected by the interventions under study provided an additional motivation for a prescribed age-at-enrollment distribution. Table 1 provides information on enrollment by age group in the various WHI components.

In addition to the 40 participating clinical centers, the WHI program is implemented through a clinical coordinating center based at the Fred Hutchinson Cancer Research Center in Seattle. Several components of the National Institutes of Health (National Heart, Lung and Blood Institute, National Cancer Institute, National Institute of Aging, National Institute of Arthritis, Musculoskeletal and Skin Diseases, NIH Office of Women's Health, and NIH Director's Office) sponsor the WHI program, with NHLBI taking a coordinating role.

Several important statistical issues have arisen in the design, conduct, and analysis of the WHI. Some of these, where additional methodology developments are required, will be described below in some detail.

2. Study Design

Most aspects of the CT and OS design, including target sample sizes, eligibility criteria, primary and secondary clinical outcomes, biological specimen collection and storage protocols, quality-assurance procedures, and CT monitoring and reporting methods, have previously been described (Freedman et al., 1996; Women's Health Initiative Study Group, 1998; Anderson et al., 2003; Prentice and Anderson, 2005). There are, however, study design issues related to the nutritional and physical activity epidemiology goals of the program, as well as design issues related to the efficient uses of the WHI specimen repository for genomic and proteomic purposes, that remain under active consideration.

2.1 *Nutritional and Physical Activity Epidemiology*

The reliable assessment of nutrient consumption and activity-related energy expenditure constitutes central challenges in nutritional and physical activity epidemiology. In fact, a principal argument in support of the need for the DM trial of a low-fat eating pattern, and for the CaD trial, as opposed to a reliance on observational study designs, comes from dietary assessment uncertainties and their potentially dominant impact on nutritional epidemiology association studies. Very similar measurement issues arise in physical activity assessment as most nutritional and physical activity association studies rely on self-report assessment methods. Of particular current interest are dietary and physical activity

Table 1
Women's Health Initiative sample sizes (% of total) by age group

			Postmenopausal hormone therapy							
Age group	Dietary modification		Without uterus (E-alone)		With uterus (E+P)		Calcium and vitamin D		Observational study	
50–54	6,961	(14)	1,396	(13)	2,029	(12)	5,157	(14)	12,386	(13)
55–59	11,043	(23)	1,916	(18)	3,492	(21)	8,265	(23)	17,321	(18)
60–69	22,713	(47)	4,852	(45)	7,512	(45)	16,520	(46)	41,196	(44)
70–79	8,118	(17)	2,575	(24)	3,575	(22)	6,340	(17)	22,773	(24)
Total	48,835		10,739		16,608		36,282		93,676	

patterns that may be associated with long-term energy balance in view of the obesity epidemic in North America and other Western countries, and the strong association between obesity and such major chronic diseases as diabetes, CHD, and cancer (e.g., Calle et al., 2003). A recent commentary (Prentice et al., 2004) focused on the future research agenda in the nutrition, physical activity, and chronic disease areas, and pointed to nutrition and physical activity assessment and modeling as key areas for further methodologic and substantive research.

The validity of the intervention versus control group comparisons in the DM trial does not rely directly on dietary assessment among participating women. Indeed, this lack of reliance, along with the absence of confounding by baseline risk factors, is the major motivation for an intervention trial. Dietary assessment, however, is needed for the evaluation of adherence to nutritional goals, and for explanatory analyses that attempt to attribute intervention effects on clinical outcomes to specific nutritional changes (e.g., reduced total fat, increased fruits and vegetables) induced by a multifaceted intervention program. Of course, WHI CT and OS data will be used to examine many nutritional and physical activity epidemiology associations beyond those tested by CT interventions. For these other association analyses, nutritional and physical activity assessment data will play a direct and central role.

Diet and physical activity are typically assessed in epidemiologic studies using frequencies, records, or recalls. For example, a food-frequency questionnaire (FFQ) or an activity-frequency questionnaire provide a list of foods or activities and ask a respondent to specify how frequently each is consumed or engaged in, and with what portion size or intensity, over the preceding few months. It has long been known from reliability studies (e.g., Willett et al., 1985) that these types of assessment procedures may incorporate substantial random measurement error, but evidence is emerging from biomarker studies concerning the presence of important systematic measurement error as well (e.g., Heitmann and Lissner, 1995; Day et al., 2001; Kipnis et al., 2003; Subar et al. 2003; Hebert et al., 2004). Systematic bias may occur when a person consistently tends to under- or overreport the consumption of certain foods, or the practice of certain activity patterns on successive application of the same or different self-report instruments. Relaxing the classical measurement error model (e.g., Carroll, Ruppert, and Stefanski, 1995) to include an independent person-specific random effect may help to deal with the resulting correlated measurement errors, but this modeling device will be insufficient if the systematic component to the measurement error tends to depend on individual characteristics, such as body mass, ethnicity, age, or social desirability factors. Instead, the measurement model may be conditioned on a vector, $\mathbf{V}$, of such characteristics, with the mean and variance of a random effect allowed to depend on $\mathbf{V}$.

These self-report measurement issues may cause one to instead consider biomarkers that plausibly adhere to a classical measurement model for nutritional or physical activity assessment. In fact, suitable biomarkers are available for short-term total and activity-related energy expenditure (Schoeller et al., 2002), and for protein, sodium, and potassium consumption (Bingham et al., 2002) among weight-stable persons, through a doubly labeled water protocol, urinary recovery, and indirect calorimetry. However, some of these measures (e.g., energy expenditure using the doubly labeled water technique) are quite expensive and practical only in a moderate-sized subset of an epidemiologic cohort. Hence, the viable research strategy to reliable epidemiologic association analysis seems to be to carry out a classical measurement error biomarker substudy in a suitable subset of a study cohort, and use this substudy to calibrate the self-report data that are available for the entire study cohort. For example, Prentice et al. (2002) consider a model

$$X = Z + \varepsilon \tag{1}$$

for a nutrient consumption or activity-related energy expenditure measure Z having biomarker measure X, where the error variate ε is independent of Z and other study subject characteristics (V), and the variance of ε is estimated using a repeat application of the biomarker protocol in a reliability subsample. The corresponding model for a self-report assessment, W, of Z was modeled as

$$W = \alpha + \beta Z + \gamma^T \mathbf{V} + \delta^T Z \otimes \mathbf{V} + U + e, \tag{2}$$

where, again, $\mathbf{V}$ is a vector of study-subject characteristics that may relate to the self-report measurement properties, while U is a mean zero random effect for the study subject that allows repeat assessments W to be correlated (given $\mathbf{V}$) and e is an independent error term. Some development of logistic regression estimation procedures to relate a disease odds ratio to the underlying nutrient or activity exposure Z under this measurement model, using regression calibration, conditional scores, and nonparametric corrected scores procedures (e.g., Carroll et al., 1995; Huang and Wang, 2000), is included in an unpublished 2003 Department of Statistics, University of Washington doctoral dissertation by Elizabeth Sugar.

Study design issues related to the use of models (1) and (2), or variations thereof, arise from the need to specify a sample size and sampling procedure for a biomarker subsample. Related issues concern the selection of reliability subsamples for both X and W. Suitable design choices, under (1) and (2), likely relate strongly to the relative magnitudes of the variances of ε, U, e in relation to the variance of Z, and to the dependence of such variances on $\mathbf{V}$, and also to the magnitude of the regression coefficients in (2), particularly β and δ. There are, of course, related analysis issues concerning consistent and efficient means of estimated odds ratios or hazard ratios for clinical outcomes of interest, the robustness of such inferences to moderate departures from (1) to (2), and the choice between (1) and (2) and other measurement error models.

At the time of this writing, a Nutrient Biomarker Study among 543 women in the DM component of the Women's Health Initiative CT (50% control, 50% intervention) was just being completed with a principal goal of elucidating trial results in terms of the components of this multifaceted intervention through a biomarker calibration of FFQ data. A grant proposal to study the comparative measurement properties of the FFQ, a 4-day food record and (three) 24-hour recalls, and to study the comparative properties of an activity frequency questionnaire, a 7-day physical activity recall, and

WHI personal habits questionnaire, among 450 OS women is also pending. These efforts not only include the "recovery" biomarkers (Kaaks et al., 2002) listed above, but also blood serum concentration measures for various nutrients. The classical measurement model (1) will typically be implausible for these concentration markers, so additional design and analysis issues arise in attempts to use these biomarkers in conjunction with self-report assessments in nutritional and physical activity–disease association analyses.

Since few full-scale dietary intervention trials with clinical outcomes are practical at any point in time for reasons of cost and logistics, these measurement error modeling and analysis activities become key to progress in these important population science research areas.

2.2 *High-Dimensional Genomic and Proteomic Studies*

The WHI includes a well-developed system for the standardized collection and storage of biological materials from participating women. This includes the storage of blood plasma and serum, as well as white blood cells for DNA extraction. These specimens in the well-characterized CT and OS cohorts, with comprehensive outcome ascertainment, provide an extremely valuable resource for elucidating mechanisms that determine chronic disease risk, and for explaining CT intervention effects. The WHI includes a substantial number of externally funded ancillary studies, as well as a few internally funded case–control studies, that make use of these specimens. Ideas for priority uses of specimens include high-dimensional approaches to studying genotype, or to studying serum protein expression patterns, or changes in such patterns over time. The technological advances that allow genome-wide scans of hundreds of thousands of single nucleotide polymorphisms (SNPs), from a minute amount of DNA, are impressive indeed. Though the technology is less mature, there are also several platforms for high-dimensional proteomics. However, suitable statistical methods for the design and analysis of case–control studies that include such high-dimensional data are essential for these innovations to have their desired impact on medicine and public health, and much related statistical work remains to be carried out (e.g., Feng, Prentice, and Srivastava, 2004).

Consider genetic association studies which examine the relationship of genotype to disease risk. Genotype can be characterized using the several million SNPs (Kruglyak, 1999) that exist in the human genome. There is substantial effort, including the publicly funded HapMap project, to identify a reduced set of tag SNPs that convey most genotype information as a result of correlation (linkage disequilibrium) between neighboring SNPs (Gabriel et al., 2002; Gibbs et al., 2003). Use of "chip" technologies has allowed genotyping costs to fall to the vicinity of $0.01 per SNP and certain organizations make 50,000–250,000 tag SNPs commercially available, the latter number having potential to characterize most of the common variability across the human genome. Furthermore, SNP determinations are evidently quite accurate and can be based on amplified DNA, so that as little as 1 *mcg* of DNA is sufficient for a rather comprehensive genome-wide scan.

However, large numbers of cases and controls are needed to detect associations of plausible magnitude between a given SNP and disease risk for such complex diseases as cardiovascular diseases and cancers, especially when such association is dependent on linkage disequilibrium that is less than one due to the use of tag SNPs. For example, to detect an odds ratio of 1.5 for the presence of one or both copies of the minor allele of an SNP having an allele frequency of 0.1 at the 0.05 level of significance, one would require 763 cases and 763 controls for 80% power, and 1301 cases and controls for 95% power (e.g., Breslow and Day, 1987). At 1 cent per SNP, a study of 250,000 SNPs in 1000 cases and 1000 controls would involve genotyping costs of $5 million, and would be expected to yield 12,500 "false positive" associations under the global null hypothesis of no SNP–disease associations. This implies the need for a larger sample size, or a multistage design to screen out most of the false positives, and argues for additional innovation to reduce genotyping costs.

One approach to reduce genotyping costs is to restrict the analysis to the subset of SNPs that are within the coding or regulatory regions of known genes. This is a logical and attractive approach, though there is considerable debate about the potential biologic importance of polymorphisms outside of these regions. A second interesting approach involves the pooling of equal amounts of DNA from each case (or control) prior to genotyping. Though the concept of genotyping from pooled DNA has existed for some time, much of the pertinent literature is quite recent (see Sham et al., 2002 for a review). Recent studies (e.g., Le Hellard et al., 2002; Mohlke et al., 2002) document the agreement that can be achieved between allele frequency estimates from pooled DNA compared to individual SNP genotyping. Some additional variation is introduced by using an allele frequency estimate for the set of cases (or controls), rather than an allele frequency measurement, though this additional variation can be controlled by employing a small number of replicate pools, and/or by drawing replicate samples from each pool. For example, if one formed two case pools and two control pools, each of size 500, carried out four polymerase chain reaction (PCR) amplifications from each, and quadruplicate sampled from each PCR pool, one would incur $160,000 genotyping costs for 250,000 SNPs at 1 cent/SNP. This represents a 30-fold cost reduction relative to corresponding individual genotyping, evidently with little reduction in power (Mohlke et al., 2002) for determining SNP–disease associations. This cost reduction factor is somewhat optimistic in view of pool formation costs, and necessary specialized whole genome DNA amplification procedures, but the use of an initial pooled DNA step may often be essential for an epidemiologic study to be practical in terms of cost.

A limitation of the pooled DNA approach is that one is unable to examine the joint association with disease risk of adjacent SNPs (haplotypes), or SNP–SNP interactions more generally, from pooled DNA, so there are important research strategy trade-offs to consider. Multistage study designs that employ pooling at the early stages in an attempt to screen out many of the false positives, followed by individual genotyping stages, may have considerable appeal in some settings, and deserve formal evaluation of statistical properties. Other statistical design issues relate to preferred pool sizes with some researchers evidently advocating smaller pool sizes (Barratt et al., 2002; Downes et al., 2004) than do others (Le Hellard et al., 2002; Mohlke et al., 2002) based on components of variance considerations.

A referee has pointed out that the use of pooled DNA at a given study design stage will also preclude the study of the SNPs tested in relation to other traits (e.g., hypertension) for which data may be available for individuals in the cohort, unless such trait values were specifically used in pool construction.

A multistage design seems attractive in this high-dimensional setting, whether or not pooling is employed, for reasons of excess cost and false-positive avoidance. For example, with 250,000 SNPs a three-stage design with equal sample sizes at each stage could be carried out by testing at the 0.022 level ($Z = 2.30$) at each stage, giving an expected 2.5 false positives overall under the global null hypothesis. This design would screen out nearly 98% of the SNPs at the first stage, and would involve only about 120 SNPs that are unrelated to disease at the third stage, with close to a two-thirds reduction in genotyping costs. However, further evaluation is needed of corresponding statistical properties (e.g., power properties relative to a single-stage design that tests at a very extreme significance level of 0.00001). See Sagatopan, Venkatraman, and Begg (2004) for some related encouraging power analyses.

At the time of this writing, the WHI is in the early stages of implementing a three-stage design to identify SNPs, or haplotypes, that relate to the risk of CHD, stroke, or breast cancer and to identify SNPs or haplotypes that relate to the magnitude of combined hormone (E+P) effects on these diseases. The first two stages will be in the OS, the first involving pooled DNA, while the third will take place in the E+P trial cohort, which has the most reliable information on E+P effects.

The relationship between serum (or plasma) protein concentrations and disease risk has great potential for the early detection of disease, and for the study of disease processes and intervention mechanisms. Equally important, changes in high-dimensional serum protein patterns as a result of treatment or intervention activities have great potential for preventive intervention development and initial screening, as knowledge develops on the associations of such patterns with a range of clinical outcomes. This seems fundamental as preventive intervention development to date has needed to rely on extrapolations from therapeutic trials and on low-dimensional intermediate outcome trials, both of which may lack sensitivity, or on observational epidemiology, which may often lack specificity.

Mass spectrum profiles provide an estimate of protein (peptide) intensity as a function of the peptide mass to charge ratio. Serum specimens, and hence these profiles, are, however, quite sensitive to specimen handling and processing methods, and measurement platforms differ in their resolution and other measurement properties. A multistage sequential design (Feng et al., 2004) is attractive also in this context for the identification of peptide peaks that distinguish cases from controls. Such peaks can then be studied in more detail to identify the distinguishing peptides and proteins. These analyses are more greedy in terms of specimen usage, so that a multistage design could allow poorer quality specimens to be used at the early stages (with false positives due to specimen collection or processing differences screened out at later stages) saving the better quality specimens (e.g., prediagnostic specimens collected under a standardized protocol in a cohort study or intervention trial) for the final design stages. Additional proteomic platforms that fractionate proteins according to additional features, such as affinity tags or elution times, are under vigorous development, and some are suitable for high-throughput applications, or will be in the near future.

These genomic and proteomic design issues, and associated high-dimensional data analysis issues (e.g., Tibshirani and Efron, 2002; Simon et al., 2003; Diamandis, 2004), deserve the attention of the statistical community in the upcoming years, and are expected to be crucial to the longer-term productivity of the WHI.

3. CT Monitoring and Reporting Methods

Each CT component has its designated primary and secondary clinical outcomes, and in the case of the two HT trials a designated primary adverse outcome (breast cancer). The CT monitoring guidelines, adopted by the external Data and Safety Monitoring Board (DSMB) comprised of senior researchers and clinicians having expertise in relevant areas of medicine, epidemiology, nutrition, biostatistics, CTs, and ethics, included a special role for the designated primary outcome(s). This primary outcome was CHD for the HT trials, breast cancer and colorectal cancer separately for the dietary modification trial, and hip fractures for the CaD trial.

It was also recognized from the outset that the interventions under study had potential to affect the risk, either beneficially or adversely, for various clinical outcomes beyond the primary outcome(s), and that these other effects should enter early trial stopping considerations. Hence for the HT trials the monitoring plan involved reviewing weighted log-rank statistics for breast cancer, stroke, pulmonary embolism, hip fractures, colorectal cancer, endometrial cancer (E+P trial), and deaths from other causes, in addition to CHD. For the DM trial, weighted log-rank statistics were reviewed for CHD, and deaths from other causes in addition to breast and colorectal cancer, while for the CaD trial colorectal cancer, breast cancer, fractures other than hip, and deaths from other causes were reviewed, in addition to hip fracture. The weights were linear from zero at randomization up to a plateau point at 3 years for cardiovascular disease and fracture incidence, and at 10 years for cancer and mortality. These weights were chosen to enhance the power of outcomes comparison between randomization groups, under the hypothesized time course of intervention effects. These weights were not well suited to the identification of any early adverse effects, a fundamental element of data and safety monitoring, so that unweighted log-rank statistics and Cox model hazard ratio estimates and confidence intervals were also routinely provided to the DSMB in biannual CT monitoring reports.

An important statistical and substantive issue concerns the means of usefully summarizing the benefits and risks of an intervention that may plausibly affect multiple clinical outcomes, each with its own time course, incidence rate pattern, and severity. Following a series of exercises in which DSMB members individually specified their recommended course of action concerning trial continuation (stop, continue, do not know) under scenarios as to how the data may look at a future point in time (Freedman et al., 1996) a so-called global index was developed as a part of the CT monitoring procedure. For each CT component, the global index was

defined for each participating woman as the time to the first occurrence of the clinical outcomes listed in the preceding paragraph, each of which was regarded as a major health event. If the primary outcome for a CT component, or the primary adverse outcome for the HT trials, showed significant difference between randomization groups, the global index was to be examined with early stoppage considerations for benefit or risk based on weighted log-rank statistics for the global index. The DSMB agreed to pay attention to these monitoring statistics, but not necessarily to be bound by them, and the DSMB also viewed data on a number of additional clinical and behavioral outcomes as a part of their overall assessment and safety monitoring activities.

While available statistical methods for the analysis of correlated failure times (e.g., Kalbfleisch and Prentice, 2002, Chapter 10) mostly focus on analyses of marginal hazard rates, the WHI CT highlights the importance of carefully selected summary measures of treatment effect that can guide the monitoring and interpretation of CT data. The global index defined above did play an influential role in the early stoppage of the combined hormone trial (Writing Group for the Women's Health Initiative, 2002) when the DSMB judged that risks exceeded benefits over a 5-year usage period, and has been the subject of some discussion and debate ever since. Some critics have asked, for example, why hip fracture was included but not vertebral or other fractures. No doubt there is no uniquely suited single index in such a complex setting, and additional calculations to examine the sensitivity of conclusions to inclusion and exclusion choices, and to the specification of weights among various outcomes, may be a useful element of data presentation and summary. On the other hand, however, the absence of an attempt to specify pertinent summary measures in advance of the outcome data coming available leaves an undue likelihood that post hoc debate would too strongly influence trial interpretation and clinical practice and public health impact.

The estrogen-alone CT component also was stopped early (Steering Committee for the Women's Health Initiative, 2004). In the reporting of principal results from the two HT trials, we presented hazard ratio estimates, as well as nominal and adjusted confidence intervals. The adjusted confidence intervals accommodated the sequential data examination of evolving data using an O'Brien–Fleming approach, while the elements of the global index other than the primary outcome (and primary adverse outcome) were also adjusted according to the number of elements of the global index, using a Bonferroni procedure. These latter intervals were substantially conservative since most outcomes in the global index were expected to have only a small influence on early stopping, and the Bonferroni emphasis on controlling experiment-wise error is not so natural in this setting. On the other hand, the nominal intervals are somewhat liberal, especially for the primary outcomes that may have greater influence on early stopping. Some critics of the combined hormone trial results have been quick to adopt the conservative adjusted intervals and declare some differences, where nominal but not adjusted confidence intervals excluded one, as "not significant." It would be useful to have further development of statistical monitoring and reporting methods that would lead to more specifically suited tests and confidence intervals in these types of complex situations.

4. The Roles of Clinical Trials and Observational Studies in Population Science Research

A major issue in the chronic disease prevention and population science research area concerns the designs that are needed to obtain reliable information on disease associations and intervention effects. Large-scale observational studies, especially cohort studies, allow study of the associations between a wide variety of exposures or characteristics and clinical outcomes of interest. Controlled intervention trials on the other hand represent the gold standard for studying the effects of a given treatment or intervention, in spite of typically high costs and demanding logistics. Clearly, rather few full-scale intervention trials with disease outcomes can be afforded, so the question is better focused on the interplay and complementary role that can be fulfilled by the two study designs. Hence, pertinent questions relate to the criteria, and the hypothesis and intervention development processes, that are needed to establish the feasibility and potential of a full-scale intervention trial.

4.1 *Combined HT and Cardiovascular Disease*

The rather few situations where there is evidence from observational studies and from one or more intervention trials provide an important opportunity to examine this interplay. The WHI HT trials and a large body of preceding observational studies provide such an opportunity. In fact, few research reports have stimulated as much public response (The End of the Age of Estrogen, 2002; The Truth about Hormones, 2002) or have engendered as sustained a discussion among medical practitioners and researchers as the results of the WHI E+P. While a major reduction in CHD incidence had been hypothesized based on a substantial body of observational research (Stampfer et al., 1991; Grady et al., 1992; Barrett-Conner and Grady, 1998), the WHI E+P trial found an elevation in CHD risk, and assessed that overall health risks exceeded benefits over an average 5.6-year follow-up period (Writing Group for the Women's Health Initiative, 2002; Manson et al., 2003). Table 2 shows Cox model hazard ratio estimates and nominal 95% confidence intervals from the E+P trial, and from the companion E-alone trial, from the Writing Group for the WHI (2002) and WHI Steering Committee (2004), respectively, where confidence intervals adjusted for multiple testing can also be found. Note the apparent impact of E+P, and to a lesser extent E-alone, on multiple important clinical outcomes.

The lack of explanation for the departure of E+P trial results on CHD, from expectation based on observational studies, has prompted some clinicians and researchers to hypothesize flaws in the WHI trial (e.g., Creasman et al., 2003; Goodman, Goldzieher, and Ayala, 2003). Others have argued lack of relevance of trial results to important sub-groups of combined HT users. For example, a recent contribution noted that WHI was not designed to provide a powerful test of cardioprotective effects among 50- to 54-year-old women in menopausal transition, and concluded that observational studies provide "the only applicable clinical guide to this issue" (Naftolin et al., 2004).

Other authors have speculated on reasons for a discrepancy between WHI E+P trial results and related observational research citing confounding in observational studies, the limited ability of observational studies to assess

Table 2
Clinical outcomes in the WHI postmenopausal hormone therapy trials

	E+P trial		E-alone trial	
Outcomes	Hazard ratio	95% CI	Hazard ratio	95% CI
Coronary heart disease	1.29	1.02–1.63	0.91	0.75–1.12
Stroke	1.41	1.07–1.85	1.39	1.10–1.77
Venous thromboembolism	2.11	1.58–2.82	1.33	0.99–1.79
Invasive breast cancer	1.26	1.00–1.59	0.77	0.59–1.01
Colorectal cancer	0.63	0.43–0.92	1.08	0.75–1.55
Endometrial cancer	0.83	0.47–1.47	–	–
Hip fracture	0.66	0.45–0.98	0.61	0.41–0.91
Death due to other causes	0.92	0.74–1.14	1.08	0.88–1.32
Global index	1.15	1.03–1.28	1.01	0.91–1.12
Number of women	8506	8102	5310	5429
Follow-up time, mean (SD), months	62.2 (16.1)	61.2 (15.0)	81.6 (19.3)	81.9 (19.7)

short-term effects, differences among combined HT preparations, and differences among populations of women studied as possible reasons (Grodstein, Clarkson, and Manson, 2003; Michels and Manson, 2003; Ray, 2003). The April 2004 issue of the *International Journal of Epidemiology* includes several commentaries on this topic that illustrate the continuing diversity of opinion on the sources of the discrepancy, and on the clinical implications of the available evidence.

Related perspectives on study designs that are needed to obtain reliable public health information have ranged from the statement (Herrington and Howard, 2003) that "many people suspended ordinary standards of evidence concerning medical interventions and concluded that HT was the right thing to prevent heart disease in millions of postmenopausal women despite the absence of any large-scale CT quantifying its overall risk–benefit ratio" to the assertion (Whittemore and McGuire, 2003) that "the good agreement between the observational studies and the [WHI] trial on end points other than CHD confirms the utility and validity of observational studies as monitors of new preventive agents."

Recently, Prentice et al. (2005) analyzed data from the WHI combined hormone trial among 16,608 women with a uterus, and the corresponding subset of 53,054 women in the WHI observational study who were with uterus, and not using unopposed estrogen at baseline, in an attempt to resolve this apparent discrepancy. See Langer et al. (2003) and Prentice et al. (2005) for a description of the distribution of cardiovascular disease risk factors in the two cohorts. Compared to nonusers, OS women who were using E+P preparations at baseline tended to be younger, leaner, of higher socioeconomic status, and with a lesser history of cardiovascular disease. The analyses in Prentice et al. (2005) included CHD and venous thromboembolism (VT), both of which had been shown in the CT (Writing Group for the Women's Health Initiative, 2002) to have had hazard ratios for combined hormone (E+P) use that declined with increasing time from randomization, as well as stroke. The Cox regression model

$$\lambda\{t; X(t), Z\} = \lambda_{os}(t) \exp\{x(t)'\beta_c + z\gamma\} \tag{3}$$

was employed in these analyses, where the hazard rate model for a specific clinical outcome included a λ_{os} function that was stratified (s) on baseline age in 5-year intervals, as well as cohort (CT or OS), that included treatment effects that may depend on the history $X(t)$ of E+P use up to time t following enrollment ($t = 0$) in the WHI, and baseline potential confounding factors Z. Principal interest resided in the treatment coefficients β_c, which were allowed to differ between the CT ($c = 0$) and the OS ($c = 1$). The modeled regression vector z was formed from the baseline potential confounding factors Z.

Initial analyses included an indicator variable $x(t) = 1$ if the woman was assigned to the active intervention group in the CT with $x(t) = 0$ in the placebo group, and $x(t) = 1$ if the woman was among the 33% of these OS women who were using combined hormones at baseline, and $x(t) = 0$ otherwise, without confounding factor control. For CHD, these analyses gave a hazard ratio estimate for E+P use in the OS that was only 61% of that in the CT. More specifically, the ratio (95% CI) of the E+P hazard ratio in the OS to that in the CT was 0.61 (0.46, 0.81) following simple 5-year age stratification. The corresponding ratio of hazard ratios for VT was 0.52 (0.37, 0.73), indicating that the apparent discrepancy is not just an issue for CHD. Including a vector of potential confounding factors, z, in (3) provided a partial explanation for such discrepancies as the ratio of hazard rates became 0.71 (0.52, 0.95) for CHD and 0.62 (0.43, 0.88) for VT following control for such factors as body mass index, education, cigarette smoking history, age at menopause, a baseline physical functioning measure, and age (linear) within the 5-year strata. The remainder of the discrepancy for these diseases was largely explained by acknowledging a hazard ratio dependence on time from initiation of E+P use, using the exposure history $X(t)$. In the CT, time from initiation of E+P use was defined as time from randomization with time-dependent indicator variables $x(t)' = \{x_1(t), x_2(t), x_3(t)\}$ defined according to whether women assigned to active treatment were less than 2, 2 to 5, or more than 5 years from randomization. Women using hormone therapy during screening for the hormone therapy trials were required to undergo a "wash-out" period prior to randomization. In the OS, some women had been using E+P for several years prior to enrollment. For these women, the indicator variables $x(t)$ were defined to take

Table 3
*E+P hazard ratios (95% CIs) in the CT and OS as a function of years from E+P initiation**

	Coronary heart disease		Venous thromboembolism	
Years from E+P initiation	CT HR (95% CI; $m^{\dagger}$)	OS HR (95% CI; m)	CT HR (95% CI; m)	OS HR (95% CI; m)
<2	1.68 (1.15, 2.45; 80)	1.12 (0.46, 2.74; 5)	3.10 (1.85, 5.19; 73)	2.37 (1.08, 5.19; 7)
2–5	1.25 (0.87, 1.79; 80)	1.05 (0.70, 1.58; 27)	1.89 (1.24, 2.88; 72)	1.52 (1.01, 2.29; 27)
>5	0.66 (0.36, 1.21; 28)	0.83 (0.67, 1.01; 126)	1.31 (0.64, 2.67; 22)	1.24 (0.99, 1.55; 119)

*From Prentice et al. (2005).
†m is the number of E+P group women developing disease during WHI follow-up.

value 1 according to whether the E+P usage episode prior to OS enrollment plus time from WHI enrollment was less than 2, 2 to 5, or more than 5 years at follow-up time t. A usage gap of 1 year or more defined a new hormone therapy episode.

With these definitions, and with the same potential confounding factors as in the analyses previously mentioned, there was no longer significant evidence of different treatment effect parameters between the CT and OS (Table 3) for either clinical outcome (p-values for likelihood ratio test of $\beta_0 = \beta_1$ were greater than 0.6 for CHD, and 0.8 for VT). Evidently, a major component of the apparent discrepancy for these outcomes arises from the fact that OS enrollment included few recent E+P initiators and hence little information on effects during the early years of E+P use, whereas the CT was relatively sparse following 5 or more years from randomization, while the hazard ratios decreased with increasing years from E+P initiation. The ratio of OS to CT hazard ratios for E+P (95% CI) after accounting for both years from hormone therapy initiation and confounding was 0.93 (0.64, 1.36) for CHD, and 0.84 (0.54, 1.28) for VT based on an analysis that included common β's in (3) for each of the three time periods, plus a product term between the combined hormone group indicator and the indicator for OS versus CT cohort.

Reanalyses of other observational study data, using methods like those leading to Table 3, may similarly align their results with those from the WHI E+P trial. Other factors may also prove to be important. For example, Nurses, Health Study investigators reported a substantially lower CHD risk among postmenopausal hormone therapy (E-alone and E+P) users (Grodstein et al., 2000) and this study enrolled primarily premenopausal women and hence was in a position to identify women who initiated E+P during cohort follow-up. However, apparently only biennial indicators of hormone therapy use was used in these analyses. Hence a woman who initiates E+P could be regarded as a nonuser for much of the first 2 years of use, during which the greatest hazard ratio elevation occurs. To assess the potential effects of E+P exposure data on hazard ratio estimates, we undertook an exercise in the WHI E+P trial cohort as follows. Specifically, each E+P group woman was generated a uniformly distributed ascertainment time over the first 2 years from randomization. Furthermore, we generated a random E+P stopping time. E+P group women were then regarded as nonusers up to their time of ascertainment if ascertainment preceded stopping E+P and permanently as nonusers if stopping preceded ascertainment. Motivated by hormone therapy stopping rates in community studies, the E+P stopping time density was taken to be uniform over the first 6 months with 20% stopping probability by 6 months, and uniform from 6 months to 2 years with a cumulative stopping probability of 59% at 2 years. Following final outcome adjudication, the E+P trial gave a (Manson et al., 2003) summary CHD hazard ratio (95% CI) of 1.24 (1.00, 1.54) and a standardized hazard ratio trend statistic of −2.36 (p = 0.02). This trend statistic arose by adding to the E+P group indicator variable a product term between this indicator variable and time (days) from randomization. The trend test was defined as the ratio of the maximum partial likelihood estimator for this product term divided by its estimated standard deviation. Ten runs of the contamination process just described were carried out yielding respective hazard ratio (HR) estimates (95% CI) of 1.16 (0.91, 1.47), 1.01 (0.80, 1.29), 1.25 (0.99, 1.58), 0.97 (0.76, 1.24), 1.23 (0.97, 1.55), 1.09 (0.86, 1.39), 1.13 (0.89, 1.43), 1.18 (0.93, 1.49), 1.07 (0.85, 1.36), and 1.08 (0.85, 1.37). The corresponding standardized trend statistics took values of −1.59, −1.38, −0.35, −0.07, −1.03, −2.02, −0.86, −0.59, −1.10, and −1.78. It seems evident that this type of limitation in exposure data can have important effects on study results if hazard ratios are strongly time dependent.

4.2 *Statistical Methods for Time-Varying Hazard Ratios*

Proportional hazards modeling assumptions will provide a suitable approximation in many applications. In situations where all study subjects are followed from randomization or other natural time origin for the "exposure" of interest, hazard ratio estimates arising from a proportionality assumption may provide simple and useful summary measures, even if the hazard ratio is moderately time dependent. Specifically, such estimates can be given an average hazard ratio interpretation over the study follow-up period. However, when study subjects enter a study late relative to initiation of the exposure of interest, as for hormone therapy in the OS, summary statistics calculated under a proportionality assumption may be quite sensitive to departure from a proportional hazards assumption. More generally, aspects of the hazard ratio shape may be of considerable interest in assessing the short- and long-term implications of a treatment. Statistical research is needed to develop suitable methods for summarizing treatment effects over defined exposure durations when hazard ratios are time dependent. For example, if baseline hazard rates, $\lambda_{os}(\cdot)$ in the Cox model (3), are not strongly dependent on time (t)

Table 4
E+P hazard ratios (95% CIs) as a function of years from E+P initiation, and average HRs over various times from E+P initiation, assuming common HR functions in the CT and OS

Years from E+P initiation	Coronary heart disease HR (95% CI)	Venous thromboembolism HR (95% CI)
<2	1.56 (1.12, 2.19)	2.87 (1.89, 4.35)
2–5	1.16 (0.89, 1.51)	1.70 (1.28, 2.26)
>5	0.81 (0.67, 0.99)	1.26 (1.02, 1.56)
	Average HR (95% CI)	Average HR (95% CI)
2	1.56 (1.12, 2.19)	2.87 (1.89, 4.35)
4	1.36 (1.09, 1.70)	2.28 (1.72, 3.03)
6	1.27 (1.04, 1.54)	2.07 (1.62, 2.63)
8	1.13 (0.96, 1.33)	1.83 (1.50, 2.23)
10	1.07 (0.92, 1.24)	1.71 (1.43, 2.05)

estimates of hazard ratios averaged over specified treatment durations may be useful, and can be based on estimates of β and its asymptotic distribution. For example, the upper part of Table 4 shows HR estimates for CHD and VT as a function of time from E+P initiation, when these estimates are restricted to be common to the CT and OS. The lower part of Table 4 shows corresponding average hazard ratio estimates and nominal 95% confidence, obtained using the delta method, over various time periods from E+P initiation. Note that these analyses suggest that the HR for CHD may drop below one at 5 or more years from E+P initiation. An HR below one, however, does not by itself imply cardioprotection in view of the likely selection of women at high risk for CHD at earlier times from E+P initiation. Also, the lower part of Table 4 shows an average HR estimate above one, even over a 10-year period from E+P initiation. Finally, the suggestion of an HR below one at more than 5 years from initiation derives largely from OS data, so the possibility of residual confounding needs to be kept in mind in interpreting these analyses.

More generally, one might consider ratios between treatment groups of estimates of cumulative hazards, or cumulative incidence rates, as summary measures of treatment effects in the presence of time-varying hazard functions. These measures would be more complex since estimates of baseline hazard rates would be involved. These types of summary measures could be considered for the type of step function hazard ratio model shown in Table 3, or for smooth hazard ratio models, such as that recently proposed by Yang and Prentice (2005) which includes separate parameters for short- and long-term hazard ratios with a hazard ratio function that varies smoothly with t, or for the rather general class of hazard ratio models discussed by Fahrmeir and Klinger (1998).

4.3 *Intervention Adherence and Causal Inference Methods*

The analyses described in Section 4.1 used the randomization assignment and baseline current use of hormones in the OS to define a treatment indicator variable. This was done so that we could compare hazard ratio estimates in the OS to "intention-to-treat" hazard ratio estimates in the CT, the latter having a useful interpretation and comparative freedom from assumption. The magnitude of treatment effects among persons who adhere to their treatment group assignment, however, is likely to differ from those who do not, and differential adherence patterns between the CT and OS could itself be a source of hazard ratio discrepancy. Hence, the analyses of Table 3 and the upper part of Table 4 were re-run censoring a woman's follow-up period at 6 months beyond a change in E+P group status (stopped E+P use in the active groups, or initiated hormone therapy in the control groups). As shown in Table 5, this analysis among adherent women does produce HR estimates that are somewhat more distant from unity, as expected, but the patterns are similar to those given in Tables 3 and 4. This type of adherence-adjusted analysis represents a rather simple approach to a complex issue. Other approaches (e.g., Cuzick, Edwards, and Segnan, 1997; Frangakis and Rubin, 1999) are certainly worth considering, particularly if detailed and reliable adherence histories are available. In the WHI hormone therapy trials, quantitative adherence data were obtained, primarily through the use of weighed returned pill bottles, whereas in the OS adherence data were updated through annual questionnaires, and are essentially qualitative, thereby limiting the range of adherence-adjusted analyses that can be entertained.

Table 5
Adherence sensitivity analyses of hazard ratios in the CT and OS and combined CT and OS as a function of years from E+P initiation

Years from E+P initiation	CT HR (95% CI)	OS HR (95% CI)	CT/OS HR (95% CI)
		Coronary heart disease	
<2	1.75 (1.19, 2.58)	1.03 (0.38, 2.81)	1.62 (1.14, 2.29)
2–5	1.47 (1.00, 2.17)	1.08 (0.69, 1.68)	1.28 (0.96, 1.70)
>5	0.60 (0.27, 1.29)	0.82 (0.66, 1.03)	0.81 (0.66, 1.00)
		Venous thromboembolism	
<2	3.16 (1.89, 5.31)	2.60 (1.10, 6.07)	3.01 (1.95, 4.64)
2–5	2.15 (1.37, 3.39)	1.81 (1.17, 2.81)	1.98 (1.46, 2.70)
>5	1.86 (0.87, 3.98)	1.28 (1.00, 1.64)	1.34 (1.06, 1.69)

Some authors make a strong connection between adherence-adjusted analysis and so-called causal inference (Angrist, Imbens, and Rubin, 1996) and label treatment effect parameters that would apply if there was full adherence as "causal" parameters. While it is certainly of interest to consider assumptions that would lead to identifiability of such treatment parameters, the issue of causal interpretation would seem much more closely related to the type of study design, with randomized controlled designs having a distinct advantage through the statistical independence between treatment and all baseline confounding factors, whether or not such factors can be well measured, or are even recognized. In comparison, observational study analyses typically must begin with such critical assumptions of no unmeasured confounders, an ignorable "treatment assignment mechanism," and nondifferential outcome ascertainment. These assumptions may often be uncertain enough to raise questions about the causality of any estimated associations. Adherence-adjusted analyses, whether in an observational or randomized trial setting, additionally must deal with the issues that adherence to treatment goals may be highly variable due to study subject characteristics or to properties of the intervention, and that rates of censoring of follow-up times may depend on preceding adherence histories. Hence, in realistic situations adherence-adjusted analyses are best regarded as sensitivity analyses, and associated parameter estimates (e.g., full adherence hazard ratio estimates) as data extrapolation that may be less meaningful if nonadherence arises for treatment-related reasons, but of greater interest if adherence history can be regarded as a variable intrinsic to the study subject, that is not affected by treatment.

In the WHI E+P trial it would not seem appropriate to regard adherence as an intrinsic study subject characteristic. For example, in the active treatment group a larger fraction of women than expected experienced persistent vaginal bleeding following initiation of this combined hormone regimen. The protocol called for dosage modification, or the use of other hormonal agents, in response to bleeding that persisted for several months or years, and some women chose to discontinue study pills due to this side effect. Vaginal bleeding in the placebo group was far less common, but more likely to be indicative of endometrial pathology, giving rise to biopsy and the possibility of discontinuation of study pills for other reasons. Breast tenderness was another important issue for participating women, that may be treatment related. Also, long-term adherers to treatments that have potential to affect many body organs and systems, and that are subject to high-profile media coverage, likely have many biobehavioral characteristics that distinguish them from short-term users, and it is unclear the extent to which such characteristics can be measured and adequately accommodated in data analysis. The context of a randomized controlled trial typically offers substantial advantages in providing independence between any such baseline biobehavioral factors and treatment group assignment, and also through the provision of a context for censoring rates that may depend little on such factors or upon actual adherence, provided study participants provide clinical outcome data in a comprehensive fashion regardless of their extent of adherence to intervention activities.

Issues of adherence modeling and interpretation merit continued statistical development, with much to be learned through specific applications, such as arise in the WHI.

5. Discussion

Compared to therapeutic research among persons having disease, rather few statisticians devote their energies to disease prevention research. The wide variation in the rates of chronic diseases around the world, and the results of prevention trials to date for various prominent chronic diseases (e.g., Prentice, 2004) support the concept that chronic disease risk can be impacted in a relatively few years, even at advanced ages, by practical lifestyle and pharmaceutical approaches. Statisticians have an important role to play in the realization of this potential.

There are a number of pivotal study design, conduct, and analysis issues that pose rate-limiting obstacles to progress in the primary disease prevention area. The WHI illustrates some of these, including measurement error modeling methods for the study of disease rate associations with difficult-to-measure dietary and physical activity exposures; intervention development methods using high-dimensional genomic and proteomic data; trial monitoring and analysis methods when multiple disease outcomes may be affected by an intervention; and research to elucidate the interplay between observational studies, randomized trials having intermediate outcomes, and full-scale intervention trials. Prevention research is intrinsically multidisciplinary with the statistical role at par with that of other key disciplines.

Reviewers of this article have requested additional discussion of some of the points raised above, particularly concerning the advantages and disadvantages of specifying composite indices formed by several clinical outcomes in data monitoring and analysis; concerning trial monitoring considerations for early stopping in the WHI hormone therapy trials given the possibility of hazard ratios below one after several years of use; and concerning lessons that have been learned from WHI for future clinical trial and observational study design.

While no simple index can be expected to adequately summarize intervention effects on several clinical outcomes that may each have their own time course, it seems quite important for study monitoring and reporting to specify a clear trial monitoring plan before meaningful clinical outcome data come available within the trial. In the case of each of the WHI CT components, the monitoring plan gave a special place to the trial's primary outcome, the prevention of which motivated and justified the trial, and in the case of the HT trials to an anticipated safety outcome (breast cancer). Beyond these outcomes, however, the specification of a so-called global index in an attempt to summarize benefits and risks of the intervention seemed quite valuable for trial monitoring, and the exercises (scenarios) used in developing these indices and the overall monitoring procedure were quite valuable to the DSMB. For example, these exercises facilitated the identification and resolution of differing viewpoints among board members in advance of needing to make recommendations based on trial outcome data. Of course, monitoring committees will appropriately want to examine data beyond these primary outcomes and summary indices, and the reporting of trial results could usefully include analyses of the robustness

of clinical implications to variations in the composition of summary indices, and to other aspects of the reporting process.

Some reviewers raised questions about whether the E+P trial should have stopped after an average 5.6 years of follow-up in view of the potential long-term benefits (Table 3). Certainly, these are complex and challenging decisions, and the time course of evolving and potential future risks and benefits is one of the most difficult to assimilate into trial monitoring procedures. Statistical methods for trial monitoring also seem quite limited in this respect, in that most formal sequential testing procedures make a proportional hazards assumption for outcomes that may affect an early stopping decision. In the case of the WHI E+P trial, an elevation in the designated safety outcome, breast cancer, was the trigger for an early stopping consideration under the monitoring guidelines, and this elevation was supported by a global index value indicating that risks exceeded benefits over the intervention period. These statistics were supplemented by various other less formal outcome contrasts, and conditional power calculations under various scenarios concerning future trends constituted the statistical input to early stopping considerations, with the DSMB reserving the option of making recommendations based on their own judgments which may, for example, be informed also by data external to the trial. Additional publications are under development to elaborate the data and considerations leading to the early stopping of the two WHI HT trials.

There are many lessons from WHI relative to the design of disease prevention trials and cohort studies. Two that may merit repeating relate to HR function shape in cohort study design and analysis, and the complementary role of trials and cohort studies in assessing the overall benefits and risks of a preventive intervention. If an exposure, such as hormone therapy, is a major motivation for a cohort study, then attention should be directed to the enrollment of a sufficient number of new initiators of such exposure (e.g., Ray, 2003) in order to be in a position to assess short-term intervention effects. Even if a sizeable number of new initiators are enrolled, cohort study data analyses may often need to use summary measures of exposure effect, such as average hazard ratios, to allow for time variation in hazard ratios, and to summarize exposure effects over defined exposure periods.

For reasons of cost, logistics, and ethics, preventive intervention trials may often not be able to be continued as long as would be necessary to assess risks and benefits of the long-term use of an intervention, or even to assess the longer-term risks and benefits of a relatively short-term intervention. Observational study data, strengthened by joint analysis with intervention trial data when practical, are essential for assessing such long-term effects, and for examining interactions of exposure effects with study subject characteristics, which CTs are typically not designed to do in a powerful fashion.

Finally, the surprising results from the WHI HT trials reinforce questions about the adequacy of the hypothesis development and early evaluation infrastructure for the national and international disease prevention program. Attention to observational study design and analysis issues can strengthen this infrastructure. The promise of comprehensive genomic and proteomic tools may also strengthen this "enterprise" by enhancing the development of interventions that are likely to have favorable benefit versus risk profiles, thereby setting the stage for additional valuable primary disease prevention trials.

Acknowledgements

This work was supported by grant CA-53996 from the National Cancer Institute, and by contract WH-2-2110 from the National Heart, Lung, and Blood Institute.

References

Anderson, G. L., Manson, J., Wallace, R., Lund, B., Hall, D., Davis, S., Shumaker, S., Wang, C. Y., Stein, E., and Prentice, R. L. (2003). Implementation of the Women's Health Initiative study design. *Annals of Epidemiology* **13,** 5–17.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91,** 444–455.

Barratt, B. J., Payne, F., Rance, H. E., Nutland, S., Todd, J. A., and Clayton, D. G. (2002). Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Annals of Human Genetics* **66,** 393–405.

Barrett-Conner, E. and Grady, D. (1998). Hormone replacement therapy, heart disease, and other considerations. *Annual Review of Public Health* **19,** 55–72.

Bingham, S. A. (2002). Biomarkers in nutritional epidemiology. *Public Health Nutrition* **5,** 821–827.

Bingham, S. A., Luben, R., Welch, A., Wareham, N., Khaw, K. T., and Day, N. (2003). Are imprecise methods obscuring a relationship between fat and breast cancer? *Lancet* **362,** 212–214.

Boyd, N. F., Stone, J., Vogt, K. N., Connelly, B. S., Martin, L. J., and Minkin, S. (2003). Dietary fat and breast cancer revisited: A meta-analysis of the published literature. *British Journal of Cancer* **89,** 1672–1685.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods for Cancer Research 2. The Design and Analysis of Cohort Studies*. IARC Scientific Publication 82. Lyon, France: International Agency for Research on Cancer.

Calle, E. E., Rodriquez, C., Walker-Thurmond, K., and Thun, M. J. (2003). Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *New England Journal of Medicine* **348,** 1625–1638.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. New York: Chapman and Hall.

Creasman, W. T., Hoel, D., and DiSaia, P. J. (2003). WHI: Now that the dust has settled: A commentary. *American Journal of Obstetric Gynecology* **189,** 621–626.

Cuzick, J., Edwards, R., and Segnan, N. (1997). Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in Medicine* **16,** 1017–1029.

Diamandis, E. P. (2004). Analysis of serum proteomic patterns for early cancer diagnostics: Drawing attention to potential problems. *Journal of the National Cancer Institute* **96,** 353–356.

Downes, K., Barratt, B. J., Akan, P., Bumpstead, S. J., Taylor, S. D., Clayton, D. G., and Deloukas, P. (2004). SNP allele frequency estimation in DNA pools and variance component analysis. *Biotechniques* **36,** 840–845.

The End of the Age of Estrogen [cover story]. (2002). *Newsweek* July 22.

Fahrmeir, L. and Klinger, A. (1998). A nonparametric multiplicative hazard model for event history analysis. *Biometrika* **85,** 581–592.

Feng, Z., Prentice, R. L., and Srivastava, S. (2004). Research issues and strategies for genomic and proteomic biomarker discovery and validation: A statistical perspective. *Pharmacogenomics* **5,** 709–719.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment non-compliance and subsequent missing outcomes. *Biometrika* **86,** 365–379.

Freedman, L. S., Anderson, G. L., Kipnis, V., Prentice, R. L., Wang, C. Y., Rossouw, J. R., Wittes, J., and DeMets, D. (1996). Approaches to monitoring the results of long-term disease prevention trials: Examples from the Women's Health Initiative. *Controlled Clinical Trials* **17,** 509–525.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., et al. (2003). The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., et al. (2003). The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796.

Goodman, D., Goldzieher, J., and Ayala, C. (2003). Critique of the report from the Writing Group of the WHI. *Menopausal Medicine* **10,** 1–4.

Grady, D., Rubin, S. B., Pettiti, D. B., et al. (1992). Hormone therapy to prevent disease and prolong life in postmenopausal women. *Annals of Internal Medicine* **117,** 1016–1037.

Greenwald, P. (1999). Role of dietary fat in the causation of breast cancer: Point. *Cancer Epidemiology Biomarkers and Prevention* **8,** 3–7.

Grodstein, F., Manson, J. E., Colditz, G. A., Willett, W. C., Speizer, F. E., and Stampfer, M. J. (2000). A prospective observational study of post-menopausal hormone therapy and primary presentation of cardiovascular disease. *Annals of Internal Medicine* **133,** 933–941.

Grodstein, F., Clarkson, T. B., and Manson, J. E. (2003). Understanding the divergent data on post-menopausal hormone therapy. *New England Journal of Medicine* **348,** 645–650.

Hebert, J. R., Clemow, L., Pbert, L., Ockene, I. S., and Ockene, J. K. (1995). Social desirability bias in dietary self-report may compromise the validity of dietary intake measures. *International Journal of Epidemiology* **24,** 389–398.

Heitmann, B. L. and Lissner, L. (1995). Dietary underreporting by obese individuals: Is it specific or non-specific? *British Medical Journal* **311,** 986–989.

Herrington, D. M. and Howard, T. D. (2003). From presumed benefits potential harm—Hormone therapy and heart disease. *New England Journal of Medicine* **349,** 519–521.

Huang, Y. and Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric correction approach. *Journal of the American Statistical Association* **45,** 1209–1219.

Hunter, D. J. (1999). Role of dietary fat in the causation of breast cancer: Counter-point. *Cancer Epidemiology Biomarkers and Prevention* **8,** 9–13.

Kaaks, R., Ferrari, P., Ciampi, A., Plummer, M., and Riboli, E. (2002). Uses and limitations of statistical accounting for random error correlations, in the validation of dietary questionnaire assessments. *Public Health Nutrition* **5,** 969–976.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. New York: John Wiley and Sons.

Kipnis, V., Subar, A. F., Midthune, D., et al. (2003). Structure of dietary measurement error: Results of the OPEN biomarker study. *American Journal of Epidemiology* **158,** 14–21.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22,** 139–144.

Langer, R. D., White, E., Lewis, C. E., Kotchen, J. M., Hendrix, S. L., and Trevisan, M. (2003). The Women's Health Initiative observational study: Baseline characteristics of participants and reliability of baseline measures. *Annals of Epidemiology* **13,** S107–S121.

Le Hellard, S., Ballereau, S. J., Visscher, P. M., et al. (2002). SNP genotyping on pooled DNAs: Comparison of genotyping technologies and a semi-automated method for data storage and analysis. *Nucleic Acids Research* **30,** 1–10.

Manson, J. E., Hsia, J., Johnson, K. C., et al., for the Women's Health Initiative Investigators. (2003). Estrogen plus progestin and the risk of coronary heart disease. *New England Journal of Medicine* **349,** 523–534.

Michels, K. B. and Manson, J. E. (2003). Postmenopausal hormone therapy: A reversal of fortune. *Circulation* **107,** 1830–1833.

Mohlke, K. L., Erdos, M. R., Scott, L. J., et al. (2002). High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 16928–16933.

Naftolin, F., Taylor, H. S., Karas, R., et al. (2004). The Women's Health Initiative could not have detected cardioprotective effects of starting hormone therapy during the menopausal transition. *Fertility and Sterility* **81,** 1498–1501.

Prentice, R. L. (2004). Chronic disease prevention: Public health potential and research needs. *Statistics in Medicine* **23,** 3409–3420.

Prentice, R. L. and Anderson, G. (2005). Women's Health Initiative: Statistical aspects and early results. In *Encyclopedia of Clinical Trials*, 2nd edition, P. Armitage and T. Colton (eds). New York:Wiley.

Prentice, R. L., Sugar, E., Wang, C. Y., Neuhouser, M., and Patterson, R. (2002). Research strategies and the use of nutrient biomarkers in studies of diet and chronic disease. *Public Health Nutrition* **5,** 977–984.

Prentice, R. L., Willett, W. C., Greenwald, P., et al. (2004). Nutrition and physical activity and chronic disease prevention: Research strategies and recommendations. *Journal of the National Cancer Institute* **96,** 1276–1287.

Prentice, R. L., Langer, R., Stefanick, M., et al. (2005). Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between the observational studies and the Women's Health Initiative clinical trial. *American Journal of Epidemiology* **162,** 1–11.

Ray, W. A. (2003). Evaluating medication effects outside of clinical trials: New-user designs. *American Journal of Epidemiology* **158,** 915–920.

Sagatopan, J. M., Venkatraman, E. S., and Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60,** 589–597.

Schoeller, D. A. (2002). Validation of habitual energy intake. *Public Health Nutrition* **5,** 883–888.

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA pooling: A tool for large-scale association studies. *Nature Reviews Genetics* **3,** 862–871.

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* **95,** 14–18.

Stampfer, M. and Colditz, G. (1991). Estrogen replacement therapy and coronary heart disease: A quantitative assessment of the epidemiologic evidence. *Preventive Medicine* **20,** 47–63.

Subar, A. F., Kipnis, V., Troiano, R. P., et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study. *American Journal of Epidemiology* **158,** 1–13.

Tibshirani, R. and Efron, B. (2002). Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology* **1,** Article 1, The Berkeley Electronic Press, `http://www.bepress.com/sagmb`.

The Truth about Hormones [cover story]. (2002). *Time* July 22.

Whittemore, A. S. and McGuire, V. (2003). Observational studies and randomized studies of hormone replacement therapy: What can we learn from them? *Epidemiology* **14,** 8–10.

Willett, W. C., Sampson, L., Stampfer, M. J., et al. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology* **122,** 51–65.

Women's Health Initiative Steering Committee. (2004). Effects of conjugated equine estrogen in post-menopausal women with hysterectomy: The Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* **291,** 1701–1712.

Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. *Controlled Clinical Trials* **19,** 61–109.

Writing Group for the Women's Health Initiative Investigators. (2002). Risks and benefits of estrogen plus progestin in healthy post-menopausal women. Principal results from the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* **288,** 321–333.

Yang, S. and Prentice, R. L. (2005). Semiparametric analysis of short-term and long-term relative risks with two sample survival data. *Biometrika* **92,** 1–17.

Received October 2004. *Revised February* 2005.
Accepted March 2005.

Discussions

Raymond J. Carroll

Department of Statistics
Texas A&M University
TAMU 3143, College Station
Texas 77843-3143, U.S.A.
email: *carroll@stat.tamu.edu*

Prentice, Pettinger, and Anderson are to be congratulated for an interesting and timely article.

In what follows, we will use the notation of Carroll, Ruppert, and Stefanski (1995), which is slightly different from that of Prentice et al. One of the plagues of measurement error modeling is that everyone uses the same symbols (X, W, Z, U), but their meaning is seemingly randomly permuted from author to author!

Let X denote true intake, W intake from a self-report instrument such as a food frequency questionnaire, Z study-specific characteristics, and M a biomarker. Let i denote the individual and j denote the replicated instrument. Then models such as equation (2) of Prentice et al. or the person-specific bias models of Kipnis et al. (2001, 2003) basically state that for some function $m(\bullet)$,

$$W_{ij} = m(X_i, Z_i, \mathcal{B}) + r_i + \epsilon_{ij}; \tag{1}$$

$$M_{ij} = X_i + U_{ij}, \tag{2}$$

where the random variables r_i, ϵ_{ij}, and U_{ij} are mutually independent. In most of the models in the literature, and in Prentice et al., $m(\bullet)$ is linear in true intake X, a fact that

conveniently allows identification and method of moment estimation, and later on allows one to correct risk models for the uncertainties in the self-report instrument as given in equation (1).

The random variable r_i is called a person-specific bias (Kipnis et al., 2001), indicating that two people who eat the same amount will systematically report that amount differently.

Prentice et al. briefly allude to what is probably the biggest challenge in nutritional epidemiology, which unfortunately from this statistician's perspective is not how to handle models such as (1)–(2). That issue is the difference between a recovery biomarker and a concentration biomarker. A recovery biomarker such as doubly labeled water for energy is one where the standard classical measurement error model (2) holds. When one has a recovery biomarker, the now-vast literature on measurement error modeling can be brought into play to understand design and analysis issues.

Concentration biomarkers, such as serum plasma concentrations, do not satisfy (2), but instead in their simplest form can be thought of as following

$$M_{ij} = \alpha_0 + \alpha_1 X_i + s_i + U_{ij}, \tag{3}$$

where s_i is another variance component indicating a special type of person-specific bias, namely that two people who eat the same food may process the foods differently, and systematically differ in their concentration biomarkers. One would expect the concentration biomarker person-specific bias s_i to be independent of the self-report person-specific bias r_i.

When $m(\bullet)$ in (1) is linear in X, and when $s_i \equiv 0$, it is possible to estimate the correlation between the self-report instrument W and the true intake X, a useful fact when one is setting sample sizes. However, this estimate would be sensitive to person-specific bias in the concentration biomarker. Even worse, without additional information, α_1 in (3) is not identifiable, and trying to correct relative risk estimates for measurement error then becomes problematic.

In the case of concentration biomarkers, there seem to be at least two possibilities, and we would be interested in what Prentice et al. think of them.

- The first is to abandon the idea of using measurement error methods to estimate the relative risk of X, and instead take an operational definition as in Carroll et al. (1995, Chapter 1, Section 1.5), namely to *redefine* X_i as the mythical average of M_{ij} over many replications of the concentration biomarker. In other words, redefine usual intake as measured by the concentration biomarker to be $\alpha_0 + \alpha_1 X_i + s_i$, or, more simply, to redefine the risk factor to be the concentration biomarker after removing variability in it via averaging.
- A second possibility is to do separate feeding experiments to try to understand how the concentration biomarker is related to actual intake. It is not clear whether this is feasible, and it is especially not clear whether one can get around the issue of person-specific bias in the concentration biomarker.

Acknowledgements

Research supported by a grant from the National Cancer Institute (CA-57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

References

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman & Hall CRC Press.

Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Schatzkin, A., Subar, A., and Carroll, R. J. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidermiology* **153,** 394–403.

Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R., Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003). The structure of dietary measurement error: Results of the OPEN biomarker study. *American Journal of Epidermiology* **158,** 14–21.

N. E. Day

Strageways Research Laboratory
University of Cambridge
Wort's Causeway
Cambridge CB1 8RN, UK
email: *nick.day@srl.cam.ac.uk*

Professor Prentice and his colleagues are to be congratulated on an outstanding paper. As they rightly say, the Women's Health Initiative (WHI) is perhaps the most ambitious population research investigation ever undertaken. The complexity of the interventions, the sophistication of the design, the range of endpoints for which the trial was designed to provide definitive information, together with the overall size of the trial, are deeply impressive. It is reassuring to see that the framework for the analysis is commensurate with the power of the design. The "partial factorial" design sets the standard for the design of future large-scale intervention trials, and the inclusion of an observational component has proved highly serendipitous, an aspect I will discuss later. The paper covers a range of issues, including measurement problems in nutritional epidemiology, the design of genetic studies given the technological revolution that is sweeping through the area, the reporting and monitoring of clinical trials, and the relative roles and merits of clinical trials and observational studies in population science research.

The dietary modification (DM) component of the WHI has its origins in the distant history of the WHI, and was initially the main motivation for the study. The issues are clear. Diet and nutrition, together with physical activity, appear to be key determinants of a range of major health endpoints. Diet, however, is notoriously difficult to assess accurately, a problem compounded by the fact that diet is a high-dimensional complex of factors, many of which are highly correlated. This high level of measurement error gives great uncertainty to the results of observational studies, both to the identification of the precise dietary factor of importance and the quantitative level of effect, even in fact whether there is any appreciable dietary effect. Negative results can be at least as suspect as positive ones. The hope of the WHI was that these problems could be circumvented by a randomized clinical trial. The results of the DM component of the WHI have not yet appeared, so it is too early to tell whether the optimism behind the design was justified. However, problems that were raised at the outset have not disappeared. The primary DM was to reduce intakes of total fat and saturated fat to 20% and 7%, respectively, of average daily caloric intake, while keeping total caloric intake constant. This is an intervention that is easy neither to achieve nor to maintain. The trial will, of course, be analyzed on an intention-to-treat basis, but an understanding of what the trial results mean will depend on accurate estimation of compliance over time of the intervention, and lack of change in the control arm. The intention-to-treat analysis only answers the operational question of whether this mode of delivering the intervention has an effect. The underlying question, the one of real interest, is whether sustained reduction in fat, or saturated fat, consumption modifies health outcomes. To answer this question one has to measure the degree of compliance, that is, assess fat and saturated fat intake. Prentice and his colleagues have developed more complex, and perhaps more realistic, models of the error of dietary self-assessment, together with simpler error structure models for biomarkers (models (2) and (1) in the paper). These have been used for the design of a biomarker study now under way, and which will presumably form the basis of their analysis. It is difficult to see, however, how such a biomarker study is going to resolve the issue of sustained compliance with the study protocol by both arms of the trial. First, no biomarkers are currently available either for fat or for saturated fat intake, or indeed for carbohydrate. Second, although for the so-called recovery biomarkers, at present basically total energy, protein, potassium, and sodium, model (1) may be appropriate, there is no compelling reason why model (1) would apply to blood serum concentration markers, where levels may be affected by individual endogenous or external exposure factors and the assumption of the independence of the errors may be seriously vitiated. For crucial parameters to be identifiable, some independence assumption, or equivalent, has to be made, and only for the recovery biomarkers does there appear to be compelling justification for such an assumption. It therefore seems unlikely that the self-reported fat consumption data obtained from the trial participants can be fully or credibly calibrated. However, for interpretation of the intention-to-treat analysis individual calibration is not necessary, all that is needed is an estimate of mean fat consumption on the two arms of the trial. Even these estimates of the mean, however, will prove problematic since in model (2) there is a bias term, which requires an appropriate biomarker study for its estimation. It is also, as a second-order problem, possible, even likely, that this bias term will depend on the dietary pattern, almost certainly different on the two arms of the trial given the nature of the intervention. If the study demonstrates an appreciable effect for the intervention on the incidence of breast cancer, interpretation will be uncontroversial. If, however, the breast cancer results of the DM component are negative or only marginally positive on an intention-to-treat analysis, then interpretation will be unclear. One will not know whether the intervention produced little or no effect because fat intake is unrelated to breast cancer risk, or because the intervention did not generate sufficient difference between the two arms. Shades of the Multiple Risk Factor Intervention (MRFIT) trial may hang over the results.

The issue dealt with in this article that will attract the greatest attention, along with the companion paper in the *American Journal of Epidemiology*, relates to the effect of hormone replacement therapy on the risk for cardiovascular disease, specifically the apparent discrepancy between the consistent finding from earlier observational studies of a protective effect with the clear finding of an excess risk from the randomized component of the WHI. The results published by the WHI Writing Committee in 2002, describing an increased risk of coronary heart disease among women randomized to combined estrogen–progesterone treatment (E+P) compared to controls, gave rise to extravagant review and comment in the literature. As Prentice and colleagues point out, an issue of the *International Journal of Epidemiology* was devoted to the topic, with lurid titles to papers such as "Is this the end of observational epidemiology?" Many pet theories and old hobby-horses were brought out to "explain" the discrepancy. Among these was the claim that not just socioeconomic status but the pattern of socioeconomic status and deprivation since birth was of crucial importance. Without adjustment for such a complex of variables, available in virtually no observational study, results were fundamentally unreliable. A following paper purported to demonstrate the validity of the claim by showing that adjustment for a lifetime measure of deprivation gave results close to the E+P result in the WHI, using data from a cross-sectional study with information on prevalent coronary heart disease (i.e., a medical record or self-report of a physician diagnosis). Another commentary referred to the "vindication of old epidemiological theory." In an elegant if simple reanalysis of the WHI results, Prentice and his colleagues show such commentaries to be empty rhetoric. They examine the effect of one of the most basic of epidemiological variables, time since start of exposure. In cancer epidemiology, it is fundamental to the relationship between exposure and risk, and in cancer epidemiology would be considered a routine part of an analysis of cohort studies. They compare the results from the randomized component of the WHI with the results from the observational component.

When examined by time since E+P initiation, the two sets of results are as close as random fluctuation would allow. The apparent discrepancy simply disappears. In the first two years since initiation of E+P, the risk of coronary heart disease, and particularly venous thromboembolism, is high. More than

5 years after initiation of E+P, for coronary heart disease there is a substantial protective effect. Of particular note is that over 80% of the coronary heart disease cases on E+P on the observational component occur more than 5 years after E+P initiation, whereas among women taking E+P on the randomized component of the WHI, less than 20% of cases of coronary heart disease occurred 5 years or more after initiation of treatment. The analysis in the paper provides the clearest vindication of the insistence on using incident cases of disease, and treating time since onset of exposure as a basic variable of interest. Cross-sectional studies using data on the prevalence of disease can hardly hope to make a serious contribution.

A troubling aspect of the WHI results is the importance of the early results, that is, outcomes occurring within 2 years of treatment initiation, in triggering the trial stopping rules. Notwithstanding this paper, and the companion paper in the *American Journal of Epidemiology*, the headlines generated by the incomplete analysis published in 2002 will continue to reverberate. There has been a series of trials, mainly in the United States, where early stopping has led to incomplete, even misleading, data being published. Apart from this trial, the U.S. NIH intervention study on the use of tamoxifen for the primary prevention of breast cancer is another obvious example. These trials have been stopped before they have been allowed to continue sufficiently to generate data of unambiguous value for clinical or public health decisions. The stopping rules for the WHI were complex and sophisticated, yet have led to the appearance of misleading publications. More thought needs to be given, as Prentice and his colleagues stress, to the formulation of stopping rules which provide a more helpful balance between short- and longer-term effects. Conversely, again as is pointed out in the paper, many observational studies would benefit from the inclusion of adequate person-years at risk soon after exposure starts. Observational studies and clinical trials should be complementary, the former giving information on the effects of exposure under a much wider range of conditions and doses, but susceptible to bias, the latter giving potentially more accurate estimates of effect, but under much more restrictive conditions.

David L. DeMets

University of Wisconsin–Madison
K6/446a Clinical Sciences Center
600 Highland Avenue
Madison, WI 53792-4675
email: *demets@biostat.wisc.edu*

1. Introduction

Prentice et al. (1998) describe several statistical issues that arose during the design, conduct, and analysis of the Women's Health Initiative (WHI) randomized clinical trial (RCT) and observational study (OS). Some of the issues consist of including measurement error in modeling risk for dietary and physical activity assessment, interim monitoring for multiple outcomes and multiple diseases, the high dimensionality of genomic data, and time-dependent treatment group hazard ratios.

As Prentice et al. summarize, the WHI (Women's Health Initiative Study Group, 1998) was no ordinary RCT and OS. Most trials, even very large trials, have one or two treatments being tested on a single disease for each treatment with one or two major outcomes for each treatment. The WHI was probably the largest trial ever conducted, with over 68,000 postmenopausal women participating, and the OS had over 93,000 participants. The WHI RCT had three treatments under evaluation, a low-fat dietary modification (DM), a hormone therapy (HT) consisting of estrogen and progestin (EP) for women with a uterus (Writing Group for the Women's Health Initiative Investigators, 2002) and estrogen (E) alone for women without a uterus (Women's Health Initiative Steering Committee, 2004), a third treatment consisting of calcium vitamin D (CaD) supplementation. The DM arm had both breast cancer and colon cancer as primary outcomes with coronary heart disease (CHD) as a leading second. The goal was to lower a typical 40% fat content diet to 20%. The HT component had as a primary goal the reduction of CHD and reduction of hip fractures as a secondary outcome. The risk of breast cancer was a major concern. For the CaD component, the reduction of hip fractures was the primary outcome.

From a design perspective, the WHI is a formidable challenge. There is no reason to expect that the sample size requirements should be the same for each component, and in fact they were not the same. In the DM component, almost 49,000 women were enrolled. For the HT component, 10,739 patients were enrolled in the estrogen alone study (Women's Health Initiative Steering Committee, 2004) and 16,608 were enrolled in the estrogen–progestin study (Writing Group for the Women's Health Initiative Investigators, 2002), and over 36,000 were in the CaD study. Each treatment arm was compared to a control arm, which were standard diet for the DM component and a placebo for the E, EP, and CaD treatment arms in the other three components. Furthermore, women could be eligible and elect to participate in one or more of the three components (DM, HT, or CaD). In addition, the randomized cohorts needed to be stratified to achieve racial and age targets. Recruitment was to be conducted in 40 clinical centers.

Because of these complexities, a partial factorial design was used, relying on individual design and sample size calculations for each component. The WHI assumed that the individual components would be independent of each other; that is, no interaction was expected or assumed. However, there were several other multiplicities, especially in multiple outcomes for each of the three components, especially for the HT component. In addition to CHD, hip fracture, and breast

cancer, other outcomes such as stroke and specific subtypes (e.g., ischemic and hemorrhagic) as well as outcomes related to blood clotting risks (e.g., deep vein thrombosis, pulmonary embolism) arose during the conduct of the trial. How to be sensitive to various risks but yet be prudent about the increase in false claims due to multiplicities is not clear even for the standard RCT, much less a trial of this complexity.

Another challenge is that all of the three treatment components are readily available, and a belief among many groups in the medical community and the public that these are effective treatments. Thus, the challenge of adherence to the treatment arm assigned during the conduct of the trial was substantial. Based on previous observational studies by several research groups, the use of each of the three treatment modalities was associated with a reduction in risk. While the medical community fully recognized the limitation of observational studies, the use of HT, for example, was among the most widely prescribed pharmacologic agents for women.

There are several historical lessons prior to WHI about the use of observational cohort studies to infer not just associations but causality. For example, several cohort studies demonstrated an association between serum betacarotene levels and the risk of cancer, especially lung cancer. Based on these cohort studies, three major trials of betacarotene were launched. The Alpha-Tocopherol Beta Carotene (ATBC) trial was a randomized placebo control factorial trial conducted in Finland among 26,000 heavy smokers (Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group, 1994). The CARET trial was a similar design conducted in the United States among heavy smokers and industrial workers exposed, for example, to asbestos (Omenn et al., 1994). The third trial, the Physicians Health Study (PHS), was a randomized placebo control factorial trial of aspirin and betacarotene involving over 22,000 U.S. male physicians (Hennekens et al., 1996). All the three trials used a synthetic betacarotene to increase serum levels. The ATBC, at completion, indicated an increased risk of lung cancer incidence and mortality, contrary to expectations based on the observational studies. The CARET trial terminated early with an increased risk of lung cancer incidence and mortality, the rates being nearly identical to the ATBC trial. The betacarotene component of the PHS ended with a hazard ratio of nearly unity, a population that had only a small subgroup of smokers and with little exposure to other lung cancer carcinogens. Interestingly, in the placebo arms of all three trials, the baseline levels of serum betacarotene levels were associated with an increased risk of lung cancer, confirming the association seen in earlier observational studies. Yet, modification of serum betacarotene had the opposite effect. The lesson is that observational studies identify associations and should not be taken as evidence of causality and subsequent treatment strategies.

Similar lessons were learned in identifying the association of lipid values and the risk of CHD. The Framingham Heart Study (FHS) was among the first observational studies to identify this risk factoring in the late 1950s and in early 1960s (Dawber, Meadors, and Moore, 1951). Yet, several trials were able to effectively reduce serum lipid values without any benefit in reducing CHD risk. The Coronary Drug Project (CDP) was among the first trial started in the late 1960s to demonstrate that lowering serum lipid values through agents such as clofibrate did not affect CHD reductions (Coronary Drug Project Research Group, 1975). In fact, the first successful lipid reduction with a corresponding risk in CHD mortality was almost 30 years later, using a statin, zimvistatin, in a Scandinavian trial (Scandinavian Simvistatin Survival Study, 1994).

For the HT component, the observational studies did not predict the effect of either treatment modality. The reasons for this are not clear beyond the knowledge that association is not the same as causation. One possible factor is selection bias. For the HT component, women who were taking hormones were possibly more health conscious and physically active. Thus, their CHD risk was already lower and the use of hormones to treat postmenopausal symptoms induced a correlation that was not correct. Another factor is that researchers study what they can measure but there are probably many unknown but extremely important factors involved in the increased risk of CHD.

In evaluating the failure of a low-fat diet to reduce the risk of breast and colon cancer, Prentice et al. examine the impact of measurement error in dietary assessment in assessing risk. They recognize the limitations of the observational studies that suggested the low-fat hypothesis. Dietary assessment is very challenging and full of imprecision. Food frequency questionnaires are fraught with measurement errors and also susceptible to systematic bias such as over- or underreporting, conscious or not. Prentice et al. consider a model of risk assessment which incorporates measurement error in the independent variable. Measurement error is likely to have attenuated the strength of the association but still may not fully address the causation issue. The final results of the DM component are not yet available.

2. Even Higher Dimensionality

The WHI RCT and OS studies came at a time of great change and innovation in biomedical research. The sequencing of the human genome and the advances in both genomic and proteomic research offers exciting new opportunities. The WHI leaders collected and stored biological materials from the women participating in the WHI RCT and OS studies. These data from this well-characterized cohort of women will be analyzed and explored for years. The dimensionality of the data collected is far beyond anything undertaken previously.

For both epidemiology and clinical trials, current statistical methodology is simply not adequate to meet the challenges of such high-dimensional data in very large cohorts such as the WHI RCT and OS studies. New methodology, both frequentist and Bayesian based, must be developed that addresses the dimensionality and multiplicity. In addition, the laboratory methods used to measure the biological specimens is also changing rapidly as new advances are made in both the biology and the technology. Many methods such as microarrays are full of measurement error that could be improved using some of the statistical designs for laboratory quality control. For example, current results can vary with the placement of

the material on the microarray chip from run to run and from day to day.

In addition, as Prentice et al. point out, the costs of these measurements can limit the amount of data that can be collected. Of course, with time and improved technology, the costs will come down dramatically so that the volume of data generated from the WHI cohorts will be affordable.

Nevertheless, this area should serve to be a rich area for statistical research whether the environment is laboratory, epidemiological, or clinical trial investigation. The WHI may well be a leading motivation and a beneficiary as well for such statistical methodology.

3. Trial Monitoring

As suggested by the design, the WHI is a complicated trial to monitor and conduct interim analyses for early evidence of benefit or harm. There are essentially four trials being conducted, with three treatment modalities, through the same trial infrastructure, with women participating in one or more of the components. Each treatment modality can affect more than one disease, and each disease may have one or more measurements assessing treatment effect. Finally, safety monitoring for these three treatment modalities involves a multitude of outcomes.

The NIH appointed an independent Data and Safety Monitoring Board (DSMB) consisting of experts in the different treatment modalities and diseases, as well as senior biostatisticians and ethicists. All were experienced researchers and familiar with clinical trials. Not all were experienced in trial monitoring as in a DSMB. The WHI DSMB was chartered to review the WHI accumulating data at least twice per year for evidence of early benefit or harm in any or all of the treatment modalities. The DSMB could recommend continuation, a protocol modification, or early termination if the interim data were convincing. To prepare the DSMB members, the WHI leadership prepared several scenarios and surveyed the members as to what they would recommend for the WHI RCT (Freedman et al., 1996). While none of the imagined scenarios actually occurred, the process was perhaps helpful to some members and did serve to bring together the DSMB into a functioning unit.

Standard group sequential methodology was used to monitor each major primary outcome and leading secondary outcome. Some adjustments were made for multiplicities of outcomes but not all. For the HT arm, only an upper group sequential boundary for benefit was prespecified, which turned out to be a mistake. A lower boundary for harm should have been prespecified as well, perhaps an asymmetric boundary.

The EP component was terminated early due to a convincing adverse risk of clotting problems as evidenced by increases in stroke, pulmonary embolism, and deep vein thrombosis. In addition, there was an increase in breast cancer (Writing Group for the Women's Health Initiative Investigators, 2002). The trends began to emerge and kept getting stronger while there was no apparent reduction in either mortality or CHD. Hip and other fractures had a benefit with HT, as was expected. After a few meetings, the trends became convincing and the DSMB recommended to the sponsor that the EP component should be terminated. The prespecified scenarios were not so useful at this juncture, and the group sequential boundaries were helpful but still the DSMB had to render its best scientific and ethical judgment.

The E component of the HT was also terminated early but with much greater debate among the DSMB (Women's Health Initiative Steering Committee, 2004). Here, the same risk factors for clotting problems emerged as had been the case for the EP component. Hip fractures were reduced, but there was no effect on CHD in this case as well. However, in contrast to the EP component, there was a trend for a breast cancer benefit, not harm. Thus, the mix of the issues was different. The DSMB was of a mixed mind on what should be done. When the data became convincing of the clotting problems, the DSMB view was that some change needed to be made, that continuing as is was not acceptable. In a close vote, the DSMB recommended to continue the trial but to inform the participants about the clotting risks and that the breast cancer question was not resolved. This was an agonizing recommendation, with each DSMB member being split within themselves. The split vote was taken to another ad hoc committee which affirmed the recommendation of the DSMB. The trial sponsor, the National Heart, Lung, and Blood Institute, engaged in discussions with the other NIH institutes as well as the director's office. Ultimately, the NIH determined to simply terminate the WHI E component.

A global index was created which was a combination of all the major health events. The plan was to require the global index to be consistent with the results of a primary outcome before early termination should be seriously considered. However, since the global index was a combination of outcomes that were going in different directions, the global index was not as useful as originally intended. Had the directions of the major outcomes all been in the same direction, the influence may have been greater.

No additional statistical methodology would have made DSMB recommendation either easier or faster. The issues were simply too complex and while statistics was a part of the discussion, it was not the dominating factor. Still, the challenges of monitoring multiple outcomes, not totally independent, remain and further work is warranted.

4. Changing Hazards and Changing Weights

The primary analysis of the time-to-event data used a weighted log-rank test. The weights were constructed to diminish the impact of early events or early treatment effect. The rationale for this weighting is that it would not be rational for the treatments, say, for example, HT, to have an immediate impact. Thus, a modest if any treatment effect in the early going could reduce the power of the comparison unless this period of follow-up was discounted. The challenge, however, is what the weights should be. In the WHI, the weights were linear from randomization to 3 years for cardiovascular disease and fracture and 10 years for cancer incidence and mortality. Unweighted rank tests were used for safety assessment. The challenge is what lag period to use for the weighted rank tests. Many effective treatments in cardiology, such as aspirin, statins, and beta blockers,

demonstrated an effect within 3 years. For cancer, it is assumed that the process of initiation, promotion, and progression of cancer takes time, and thus no treatment can have an effect immediately. Any early cancer incidence was a process already underway and not subject to a DM prevention strategy. However, 10 years may be too long. In any case, both weighted and unweighted analyses should probably be conducted.

The issue of changing hazard ratios over the follow-up period is not new to clinical trials but was of special interest in the WHI. As Prentice et al. point out, "hazard ratio estimates arising from a proportionality assumption may provide simple and useful summary measures even if the hazard ratio is moderately time dependent." However, the hazard ratio may be sensitive to time dependency if the participants enter late relative to the initiation of risk exposure. Estimation of downstream hazard ratios is itself challenging since the participants may represent different risk groups due to differential mortality, adherence, and follow-up. That is, the different hazard ratios may be confounded. This may not have been a major issue in the WHI but is nevertheless a concern. Clearly, more research into the sensitivity of this effect would be welcome for all clinical trials, not just the WHI.

5. Intervention Adherence and Causal Inference

Since Canner first wrote about the challenge of analysis of primary outcomes adjusting for intervention compliance, based on the Coronary Drug Project, clinical trialists have recognized the dangers of this approach (Canner, 1991). Canner and others have provided examples that demonstrate that placebo compliers may have better or worse effects than placebo noncompliers. Compliance is itself an outcome and not necessarily independent of how the participant is faring in the trial. Canner also demonstrated that using a multitude of measured covariates did not make this anomaly go away.

Several authors have tried to model treatment effect based on compliance to treatment in RCTs, and then extrapolating the treatment effect under optimum compliance. However, Albert and DeMets (1994) demonstrate that such modeling is very much dependent on the independence assumption, and results can be easily misleading when this assumption is not correct. However, for OS studies, researchers have no other choice than to model treatment effect based on the degree of intervention. This is one of the areas where RCTs and OS will differ due to adherence bias, and minimizing this bias is one of the strengths of the RCT if the analysis is strictly by intent to treat.

6. Post Mortems

Whenever the results of a trial do not turn out as expected, or are not consistent with previous observational trials, as was the case for the HT component, many individuals begin to speculate about possible flaws in the clinical trial. While perhaps some trials may have critical or fatal flaws, that is not likely to be the case in the WHI. The trial was well designed, despite its complexity, well conducted in the face of public and medical biases about the effects of the interventions being studied, and carefully analyzed.

Experience indicates that we should not expect perfect congruence between observational studies and clinical trials. Observational studies are best suited to identify possible risk factors, potentially modifiable, with the hope of risk reduction. Clinical trials are best suited to test rigorously whether modification of the risk factor in fact reduces the risk of the disease under consideration.

The biostatistician must resist from being an advocate for the treatment but rather focus on whether the analysis of both the OS and the RCT is as rigorous as possible, recognizing the inherent limits of the OS design and the analysis assumptions. Objectivity must be maintained with no interest in the direction of the outcome but rather that whatever the results, they can be defended rigorously. As soon as biostatisticians lose that objectivity and operate with a bias, they lose their professional effectiveness. The results of the HT arm of the WHI RCT are pretty clear.

Observational studies will always be a primary source for identifying risk factors, even in the new era of genomics and proteomics. Given recent concerns about drug safety, observational studies will most likely be the best method for assessing long-term safety once initial treatment effectiveness has been established.

References

Albert, J. M. and DeMets, D. L. (1994). On a model-based approach to estimating efficacy in clinical trials. *Statistics in Medicine* **13,** 2323–2335.

Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine* **330,** 1029–1035.

Canner, P. L. (1991). Covariate adjustment of treatment effects in clinical trials. *Controlled Clinical Trials* **12,** 359–366.

Coronary Drug Project Research Group. (1975). Clofibrate and niacin in coronary heart disease. *Journal of the American Medical Association* **231,** 360–381.

Dawber, T. R., Meadors, G. F., and Moore, F. E. J. (1951). Epidemiological approaches to heart disease: The Framingham Study. *American Journal of Public Health* **41,** 279–286.

Freedman, L., Anderson, G., Kipnis, V., Prentice, R., Wang, C. Y., Rossouw, J., Wittes, J., and DeMets, D. L. (1996). Approaches to monitoring results of long-term disease prevention trials: Examples from the Women's Health Initiative. *Controlled Clinical Trials* **17,** 509–525.

Hennekens, C. H., Buring, J. E., Manson, J. E., Stampfer, M., Rosner, B., Cook, N. R., Belanger, C., LaMotte, F., Gaziano, J. M., Ridker, P. M., Willett, W., and Peto, R. (1996). Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *New England Journal of Medicine* **334,** 1145–1149.

Omenn, G. S., Goodman, G., Thornquist, M., et al. (1994). The beta-carotene and retinol efficacy trial (CARET) for chemoprevention of lung cancer in high risk populations: Smokers and asbestos-exposed workers. *Cancer Research* **54**(7 suppl.), 2038s–2043s.

Scandinavian Simvistatin Survival Study. (1994). Randomized trial of cholesterol lowering in 4444 patients with coronary heart disease: Scandinavian Simvistatin Survival Study (4S). *Lancet* **344,** 1383–1389.

Women's Health Initiative Steering Committee. (2004). Effect of conjugated equine estrogen in post menopausal women with hysterectomy: The Women's Health Initiative randomized clinical trial. *Journal of the American Medical Association* **291,** 1701–1712.

Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. *Controlled Clinical Trials* **19,** 61–109.

Writing Group for the Women's Health Initiative Investigators. (1998). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* **288,** 321–333.

David A. Freedman

Department of Statistics
UC Berkeley
Berkeley, California 94720-3860, U.S.A.
email: *freedman@stat.berkeley.edu*

and

Diana B. Petitti

Kaiser Permanente Southern California
393 E. Walnut Street
Pasadena, California 91188, U.S.A.
email: *diana.b.petitti@kp.org*

We thank Ross Prentice and his colleagues for a rich and provocative paper that has generated many insights in a variety of methodological areas. We also thank our editor, Xihong Lin, for organizing this discussion. Ours is an age of specialization, and we propose to consider only the effect of hormone replacement therapy (HRT) on three cardiovascular endpoints: coronary heart disease, stroke, and venous thromboembolism.

First some background. Ideas of biological mechanism and evidence from observational epidemiology led many observers to conclude that HRT was protective, reducing cardiovascular death rates by a factor of 2 or more. According to Grodstein and Stampfer (1998, p. 211, 217),

> Consistent evidence from over 40 epidemiologic studies demonstrates that postmenopausal women who use estrogen therapy after the menopause have significantly lower rates of heart disease than women who do not take estrogen . . . the evidence clearly supports a clinically important protection against heart disease for postmenopausal women who use estrogen.

Also see Stampfer and Colditz (1991) and Grodstein et al. (1996).

Such findings profoundly influenced the practice of medicine. In the late 1990s, postmenopausal hormones were best-selling drugs worldwide. About 90 million prescriptions for HRT were issued annually in the United States, corresponding to 15 million HRT users (Hersh, Stefanick, and Stafford, 2004).

Some observers remained skeptical (see, for instance, Petitti, 1994; Posthuma, Westendorp, and Vandenbroucke, 1994; Vandenbroucke, 1995). Two large clinical trials were organized to resolve the issue—Heart Progestin/Estrogen Replacement study (HERS) and Women's Health Initiative (WHI). Prentice and his colleagues were actively involved in the design and analysis of WHI. The experiments demonstrated no benefit from HRT, and some harm: WHI was stopped early, largely due to an increased risk from breast cancer among the HRT group.

Debate continues on these issues—for instance, a different mix of hormones administered along a different time path might be beneficial. See, for example, *International Journal of Epidemiology* (2004, **33,** 441–467). However, the experiments led to another major change in medical practice. Today, HRT would rarely be prescribed to prevent cardiovascular disease.

WHI had two branches, an observational study and a randomized controlled experiment. By contrast with the experiment, the observational study—like many of the other observational studies—found a protective effect from HRT. What accounts for the discrepancy? Prentice and colleagues have two answers that we find persuasive.

1. Observational studies can be misleading. Therefore, it is important to adjust for confounding variables, including socioeconomic status. This may seem obvious. It is not. The Nurses' Health Study on HRT did not adjust for socioeconomic status (Grodstein et al., 1996; Humphrey, Chan, and Sox, 2002).

2. In many contexts, including the present one, time is a crucial variable. Treatment and disease are dynamic, not static.

When arguing these points, Prentice, Pettinger, and Anderson could be read as suggesting that—if properly analyzed—the observational study agrees with the randomized controlled experiment. We would have several questions about such an interpretation.

1. Observational data can be adjusted in a variety of ways. Without experimental data, it will be unclear which adjustments to make, or how far to go.
2. Table 3 in Prentice, Pettinger, and Anderson only shows results on coronary heart disease and thromboembolism. However, even after all the modeling is done, there remains a large disparity with respect to an important cardiovascular endpoint—stroke (Prentice et al., 2005). Prentice, Pettinger, and Anderson mention stroke, but do not discuss the difficulties created by this endpoint.
3. Prentice, Pettinger, and Anderson chose for their null hypothesis equality between the two branches of WHI. However, statistical power is limited, and the choice of null greatly influences conclusions.

Power is limited because the women in the treatment arm of the clinical trial are mainly short-term users of HRT. By contrast, in the observational study, users have been taking hormones for a long time. (According to the conventions used by Prentice and colleagues, in the observational study, exposure prior to baseline is counted.)

To illustrate how substantive conclusions may be determined by apparently innocuous technical choices, we suggest the following null hypothesis: compared to the randomized controlled experiment, the observational study underestimates the risks of HRT by a factor in the range of 1.5–3, depending on risk group and endpoint (heart disease, stroke, thromboembolism). The data seem to be at least as compatible with our null hypothesis as with the null hypothesis of equivalence. These null hypotheses have rather different implications for bias in observational epidemiology.

Bias stems from incomplete adjustment. Adjustment must be incomplete, because relevant lifestyle factors are extraordinarily difficult to identify or measure. Here is one example. In observational studies, women on HRT are "compliers": they follow a treatment regime prescribed by their doctors. But compliance—even by subjects assigned to placebo in a clinical trial—is associated with favorable outcomes. A factor of 2 for compliance bias is compatible with previous literature. Compliance is thoroughly confounded with treatment in observational studies of HRT. See Petitti (1994) and Barrett-Connor (1991) for additional discussion.

HRT comes in two forms: (1) unopposed (estrogen only) and (2) combined (estrogen plus progestin). WHI considered both forms (Tables 1 and 2 in Prentice, Pettinger, and Anderson). Modeling results are presented only for the combined form (Table 3 in Prentice, Pettinger, and Anderson). Hence our focus is on combined therapy.

We turn now to a policy issue. Although WHI is tax supported, its data are not available to us. Data from clinical trials are available only rarely, and conditions may be imposed that almost preclude independent analysis. Policies governing data dissemination need to be reconsidered, although due regard must be paid to patient confidentiality. Only by thorough scrutiny can error be avoided. Transparency is the best assurance of scientific quality. For additional discussion, see Geller et al. (2004).

We would sum up the methodological lessons as follows. Rigorous causal inferences have been made using observational data, from the time of John Snow on cholera and Ignaz Semmelweis on puerperal fever. Recent examples include the health effects of smoking, and the demonstration that cervical cancer is in part a sexually transmitted disease. Indeed, most of what we know about causation in the medical sciences comes from observational studies—because experiments are often unethical or impractical. We might even suggest that observation necessarily precedes experiment. What else could provide motivation, or help define protocols?

On the other hand, observational data need to be approached with caution. When there is a conflict between observational epidemiology and experiments—HRT not being an isolated case—we think that the experiments are the ones to watch. The gap between association and causation will not generally be bridged by proportional-hazard models, even with stratification and time-dependent exposure variables. For more discussion on the relative merits of experiment and observation, see Mill (1868, Book III, Chapters VII and X).

Prentice and his colleagues deserve our thanks for the paper, and their work on WHI.

References

Barrett-Connor, E. (1991). Postmenopausal estrogen and prevention bias. *Annals of Internal Medicine* **115,** 455–456.

Geller, N. L., Sorlie, P., Coady, S., Fleg, J., and Friedman, L. (2004). Limited access data sets from studies funded by the National Heart, Lung, and Blood Institute. *Clinical Trials* **1,** 517–524.

Grodstein, F. and Stampfer, M. J. (1998). The cardioprotective effects of estrogen. In *The Management of the Menopause*, Chapter 22, J. Studd (ed), 211–219. London: Parthenon.

Grodstein, F., Stampfer, M. J., Manson, J. E., Colditz, G. A., Willett, W. C., Rosner, B., Speizerm, F. E., and Hennekens, C. H. (1996). Post menopausal estrogen and progestin use and the risk of cardiovascular disease. *New England Journal of Medicine* **335,** 453–461.

Hersh, I. L., Stefnick, M. L., and Stafford, R. S. (2004). National use of postmenopausal hormone therapy: Annual trends and response to recent evidence. *Journal of the American Medical Association* **291,** 47–53.

Humphrey, L. L., Chan, B. K. S., and Sox, H. C. (2002). Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Annals of Internal Medicine* **137,** 273–284.

Mill, J. S. (1868). *A System of Logic, Ratiocinative and Inductive*, 7th ed. (1st ed., 1843). London: Longmans, Green, Reader, and Dyer.

Petitti, D. B. (1994). Coronary heart disease and estrogen replacement therapy: Can compliance bias explain the results of observational studies? *Annals of Epidemiology* **4,** 115–118.

Posthuma, W. F., Westendorp, R. G., and Vandenbroucke, J. P. (1994). Cardioprotective effect of hormone replacement therapy in postmenopausal women: Is the evidence biased? *British Medical Journal* **308,** 1268–1269.

Prentice, R. L., Langer, R., Stefanick, M., et al. (2005). Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *American Journal of Epidemiology* **162,** 404–414.

Stampfer, M. J. and Colditz, G. A. (1991). Estrogen replacement therapy and coronary heart disease: A quantitative assessment of the epidemiologic evidence. *Preventive Medicine* **20,** 47–63. Reprinted in the *International Journal of Epidemiology* 2004, **33,** 445–453.

Vandenbroucke, J. P. (1995). How much of the cardioprotective effect of postmenopausal estrogens is real? *Epidemiology* **6,** 207–208.

Sander Greenland

Departments of Epidemiology and Statistics
University of California, Los Angeles, CA
email: *lesdomes@ucla.edu*

The randomized component of the Women's Health Initiative (WHI) is an invaluable check on observational associations. The observational component could be equally important if it is analyzed thoroughly and imaginatively, from a variety of perspectives. Although valuable, the strategies described by Prentice, Pettinger, and Anderson (PPA) cover too narrow a range. Following standard practice, they take an underidentified problem (estimate a causal effect from observational data) and force identification via rather arbitrary constraints (encoded within their models). While everyone starts this way, the approach needs to be supplemented by more realistic uncertainty assessments, at least if the authors wish to draw defensible inferences about effects from the observational study component. There is also a multiple-comparisons problem that needs to be addressed using modern techniques. Other issues arise as well.

I was a bit amused by the comment in PPA that "in realistic situations, adherence-adjusted analyses are best regarded as sensitivity analyses." I regard any causal analysis of observational data (or a randomized trial with major compliance problems) as just a piece of a sensitivity analysis; it is the piece in which results are obtained under the particular assumptions of that analysis. Because we never know that all the assumptions are correct (and in fact would wisely doubt them), we had better try more than one type of analysis. By seeing how results change as we vary our approach, we are doing a sensitivity analysis. If this variation in method is too broad, going beyond credible assumptions, we may inappropriately discount our results; conversely (and far more often), if this method variation is insufficiently broad, we may miss important sensitivities and become overconfident (Greenland, 1998). Given the potential contribution of the WHI, it seems that the planned method variation outlined by PPA is insufficient. I will suggest a few of many possible expansions. Perhaps more has been done or is planned for the analysis than PPA outlined, but in any case I should hope they address the following concerns.

1. The Need to Go beyond Hazard Ratios

One concern is the exclusive focus on hazard ratios in PPA. As a large cohort study, the WHI provides an uncommon opportunity to assess outcomes on an absolute-risk scale and on a time-to-event (years of life lost) scale. These scales can be far more relevant to decision making (both individual and administrative) than hazard ratios. A hazard ratio of 2 means something very different in terms of risk and benefits if the baseline risk is 1/100,000 versus 1/100. The difference is about 1 excess case versus 1,000 excess cases per 100,000 exposed, which is a 1,000-fold difference in health-care costs, and also a large difference in the (healthy) years of life lost. There is no clue in the tables of PPA what sort of base rates or case numbers the hazard ratios apply to, and so those results are unintelligible in absolute terms.

Even for the purposes of understanding the basic biology and biases, ratio comparisons can become obscure, especially when there is no biologic basis for assuming homogeneity of ratios across covariates. For example, PPA suggest that the inclusion of the covariate main effects $z\gamma$ in their model (3) partially explains the discrepancy between OS and CT. How much explanation would be achieved by including treatment-covariate product terms in the model? Perhaps a complete explanation remains possible by allowing for more than just time variation in the ratios.

2. Limitations of Biomarkers

PPA focus on the use of biomarkers to calibrate certain short-term measures of intake and activity. This is laudable, but has limitations for the questions that ultimately motivate funding and public interest in such research, such as "what should I eat to minimize my risk of breast cancer?" and "what dietary guidelines should we promote?" One concern is that biomarkers are not good surrogates for the treatment variables (long-term dietary intakes) in these questions; no matter how well measured, long-term biomarkers (such as hair and nail contents) are affected by many poorly understood and mostly unmeasured vagaries of individual metabolism and exposures,

while the short-term biomarkers discussed by PPA reflect current diet and behavior.

Any disconnect between actual long-term diet or behavior and its biomarker is error in the biomarker for the diet. Hence, the comparison of measured diet and biomarker is a comparison of two very noisy measures of long-term diet (with presumably independent but unknown and very differently distributed error). Models (1) and (2) in PPA appear to relate short-term measures; even if the errors in these equations are zero, the results tell us nothing about the error due to dietary variation, and it is not clear from PPA how this error will be accounted for. In any case, one must turn to long-term repeat-questionnaire data to address that variation with all its sources of error as a measure of long-term intake and behavior. Addressing these sources of error will require general uncertainty assessments, as discussed below.

3. The Need for Empirical Bayes

Turning to issues of multiplicity and screening of genetic associations, it seems very odd to me that, in 2005, anyone could neglect use of empirical-Bayes (EB) and related hierarchical procedures. The landmark work of Efron and Morris (1975) on these methods included an epidemiologic application, and today Efron and others continue to advance these approaches into very genetic problems that PPA discuss (e.g., Efron, 2004). Empirical-Bayes methodology is now textbook material, and theoretical, simulation, and case studies leave little doubt about the advantages of such techniques in multiple-inference problems (Carlin and Louis, 2000).

I have strongly advised that related random-coefficient methods be used for examining effects of multiple nutrients and other factors with hierarchical measurement structure (Greenland, 2000) as will be found in some of the WHI data. Note especially that measurement errors in nutrient intakes computed from questionnaires are compounds of at least two sources: those in questionnaire response and those in the diet-nutrient table as it applies to the foods actually eaten by the subjects (as opposed to those used to construct the table).

Another aspect of the WHI for which empirical-Bayes methods could be important is in examination of potential variation in effects across subgroups, as required for making recommendations and generalizations beyond the WHI cohorts. It has already been noted that the WHI is not representative of all targets. Even if it were, however, public-health, and clinical/personal decisions are more accurately guided by differences in risks and life expectancies for multiple outcomes than by summaries across disparate groups and outcomes (Greenland, 2005a). Providing such guidance is a problem in highly multivariate prediction, for which again empirical-Bayes methods have proved their worth.

4. Bayesian and Monte Carlo Uncertainty Assessment

The more general neglect of Bayesian approaches in PPA is regrettable, as priors are needed to achieve identification of causal effects from observational data, and it is clear that PPA have priors and use them in their analyses. For example, in their analysis of E+P stopping times in the first 2 years, PPA generate times from a peculiarly rough two-step density "motivated by hormone therapy stopping rates in community studies." Setting aside the unrealistic density, their approach here much resembles the sort of Monte Carlo sensitivity analyses (MCSA) that have recently made their way from risk assessment to epidemiology, and which closely parallel Bayesian risk assessment in their use of priors (see Greenland, 2001, 2003, 2005b for reviews and examples). I believe these methods are worth deploying to examine other sources of uncertainty in the WHI, such as residual measurement error, selection effects, and confounding. Such methods may be especially relevant for addressing the potential impact of measurement error in variables that lack validation and reliability data, including confounders such as smoking history.

5. To Summarize

The Women's Health Initiative is a remarkable achievement, providing a much-needed resource for checking and challenging results of epidemiologic studies, and it will no doubt provide new leads of its own. While I think PPA have done a good job of planning the analysis within their areas of focus, a broader strategy is needed in both the choice of outcome measures and in approaches to multiple inference and uncertainty assessment. The WHI is too valuable a resource to underanalyze.

References

Carlin, B. and Louis, T. A. (2000). *Bayes and Empirical-Bayes Methods for Data Analysis*, 2nd edition. New York: Chapman and Hall.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* **99,** 96–104.

Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalization. *Journal of the American Statistical Association* **70,** 311–319.

Greenland, S. (1998). The sensitivity of a sensitivity analysis. In *1997 Proceedings of the Biometrics Section*, 19–21. Alexandria, VA: American Statistical Association.

Greenland, S. (2000). When should epidemiologic regressions use random coefficients? *Biometrics* **56,** 915–921.

Greenland, S. (2001). Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Analysis* **21,** 579–583.

Greenland, S. (2003). The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *Journal of the American Statistical Association* **98,** 47–54.

Greenland, S. (2005a). Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). *Emerging Themes in Epidemiology* **2,** 1–4.

Greenland, S. (2005b). Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A* **168,** 267–308.

Miguel A. Hernán,[1] James M. Robins,[1,2] and Luis A. García Rodríguez[3]

[1] *Department of Epidemiology*
Harvard School of Public Health
Boston, Massachusetts 02115, U.S.A.
email: *miguel_hernan@post.harvard.edu*
[2] *Department of Biostatistics*
Harvard School of Public Health
Boston, Massachusetts, U.S.A.
[3] *CEIFE–Spanish Center of*
Pharmacoepidemiologic Research
Madrid, Spain

1. Introduction

We thank Xihong Lin for the opportunity to discuss Ross Prentice and collaborators' interesting paper. The Women's Health Initiative (WHI) randomized hormone trials evaluated the effect of postmenopausal hormone therapy on the risk of various diseases (WHI Study Group, 1998). In the first WHI trial, women were randomly assigned to either estrogen plus progestin or placebo. The rate of coronary heart disease (CHD) in the hormone group was 1.24 times (95% CI: 0.97, 1.60) that in the placebo group (Manson et al., 2003). This result was surprising because large observational studies had previously suggested a reduced risk of CHD among hormone users. Among the largest of these studies were the Nurses' Health Study (NHS) in the United States (Stampfer et al., 1991; Grodstein et al., 1996, 2000; Grodstein, Manson, and Stampfer, 2001) and a study based on the General Practice Research Database (GPRD) in the United Kingdom (Varas-Lorenzo et al., 2000).

We investigate possible sources of the discrepancy by reanalyzing the observational study data using an approach that mimics as closely as possible the published analyses of the WHI randomized trial. We then compare our approach with Prentice and collaborators'. Originally we had planned to provide reanalyses of both the NHS and GPRD data. Unfortunately, our reanalysis of the NHS data is not yet complete, so we report only the GPRD results. The GPRD is a research-oriented database that covers over 3 million residents in the United Kingdom. These individuals' general practitioners register health-care and medical information about their patients in a standardized manner. The registered information includes demographic data, all medical diagnoses, consultant and hospital referrals, and a record of all prescriptions issued. Practitioners generate prescriptions directly from the computer, ensuring its automatic recording. Validation studies have shown that 90% of information present in the patients' paper medical records, and 95% of newly prescribed drugs, are recorded in the database (García Rodríguez and Pérez Gutthann, 1998; Jick et al., 2003).

Several biologic and methodologic explanations for the discrepancy between the CHD results of the WHI randomized trial and the observational studies have been proposed (Grodstein, Clarkson, and Manson, 2003; Mendelsohn and Karas, 2005). We will focus this discussion on the impact of the following methodologic limitations of the observational studies (Grodstein et al., 2003):

1. Lack of comparability between women who initiated and did not initiate hormone therapy (healthy user bias or confounding by "indication")

   In the observational studies, women who started hormone therapy may not be comparable with those who did not start hormone therapy. On average, women who decide to initiate hormone therapy may have fewer risk factors for CHD than noninitiators. Under this hypothesis, initiation of hormone therapy would be associated with a lower risk of CHD even if hormone therapy itself has no preventive effect on the risk of CHD. That is, there would be confounding for the effect of treatment initiation.

   The WHI result cannot be explained by confounding for treatment initiation because therapy initiation was assigned at random, and thus initiators are on average comparable with noninitiators.
2. Lack of comparability between women who continued and discontinued hormone therapy ("noncompliance" bias)

   Even if there were no confounding for the effect of treatment initiation, participants in observational studies who stayed on hormone therapy for extended periods may be different from those who discontinued hormone therapy shortly after initiation. For example, women who stayed on therapy may be more health conscious than the others. Under this hypothesis, a longer duration of use of hormone therapy would be associated with a lower risk of CHD even if hormone therapy itself has no preventive effect on the risk of CHD. That is, there would be confounding for the effect of treatment discontinuation.

   Similarly, WHI hormone users who stayed on hormone therapy for extended periods and those who discontinued hormone therapy shortly after initiation may not be comparable because treatment discontinuation was not randomized. The nonnull WHI results, however, cannot be explained by confounding for treatment discontinuation because the analysis was conducted under the intention-to-treat (ITT) principle. That is, the effect of hormone therapy was estimated by comparing the CHD

risk of those randomly assigned to hormone therapy and placebo, regardless of whether they complied with their assigned treatment. The ITT effect will generally be closer to the null than the effect had all women fully complied with their assigned treatment.

3. Imprecise ascertainment of the time of hormone therapy initiation

   In some observational studies (e.g., the NHS), data on hormone use was collected by questionnaires mailed every 2 years and the time of therapy initiation within the 2-year interval is largely unknown. This uncertainty introduces bias in the effect estimates over any fixed (say, 2-year) interval after treatment initiation. For example, in previous analyses, women in the NHS were assigned to the hormone use group that they reported in the questionnaire returned at the onset of the 2-year interval. Thus women who initiated therapy during the interval were systematically misclassified as nonusers until the next questionnaire. If hormone therapy initiation causes a short-term increase in risk, then this misclassification would downwardly bias the effect estimate. In the WHI there is no uncertainty regarding the time of randomized therapy initiation.

In this article, we provide reanalyses of the GPRD that only suffer from limitation 1. Limitation 3 is not present in the GPRD study because exact dates of treatment initiation are recorded. We remove limitation 2 by reanalyzing the GPRD study using an ITT principle. This reanalysis requires conceptualizing the observational GPRD study as if it were a sequence of randomized trials in which the randomization probabilities are unknown. Our ITT effect estimates from the GPRD study are then compared to the ITT estimates from the WHI randomized trial.

In Section 2, we describe a study protocol for the GPRD trials that mimics as closely as possible that of the WHI trial. In Sections 3 and 4, we reanalyze the GPRD trials and obtain (i) estimates of the ITT effect of hormone therapy and (ii) estimates of the effect of continuous hormone therapy (i.e., in the absence of noncompliance). In the last section, we compare our approach with Prentice and collaborators'.

2. Study Protocol of the GPRD Trials

2.1 *Eligibility Criteria*

We defined inclusion and exclusion criteria in our GPRD trials to mimic the WHI criteria. Like the WHI trial, the GPRD trials include only women aged 50 years or more and with an intact uterus. We mimicked the WHI exclusion criteria (WHI, 1998) as closely as we could by excluding GPRD women with a past diagnosis of cancer (except nonmelanoma skin cancer), cardiovascular disease, and cerebrovascular disease (Varas-Lorenzo et al., 2000).

2.2 *Baseline and Follow-Up*

In the WHI, women were followed from the time of randomized treatment assignment (baseline) to the diagnosis of a CHD endpoint, death from causes other than CHD, loss to follow-up, or administrative end of follow-up, whichever came first.

In the GPRD cohort, we need to define the time of "randomized" treatment assignment (baseline). Because the follow-up of our cohort started in January 1991, we can define baseline as January 1991, apply the eligibility criteria to women in the cohort in January 1991, and compare the CHD risk of eligible women who reported treatment initiation with that of eligible women who did not report treatment initiation during January 1991. Alternatively, we can define the baseline as February 1991, or as any other subsequent time before the end of follow-up in December 2001. For each possible baseline time, we can apply the eligibility criteria to women in the cohort at that time so women participating in the trial starting in January 1991 would not necessarily be the same women participating in the trial starting in, say, December 1994.

But rather than fixing a single baseline month for our GPRD trial, we can conduct all possible trials, pool the data, and obtain an estimate of effect with a narrower confidence interval (which appropriately accounts for correlations that may arise from using the same individuals in several trials). Let m denote month with $m = 0, 1, \ldots, 131$ representing January 1991, February 1991, $\ldots$, December 2001. We started a separate GPRD trial at each month m. Each woman may participate in a maximum of 132 trials. For each trial, follow-up started in month m (baseline) and ended at diagnosis of a CHD endpoint, death from causes other than CHD, loss to follow-up, or administrative end of follow-up (8 years like in the WHI or December 2001), whichever came first. We index trials by the month m in which they start.

2.3 *Treatment Regimes*

WHI participants were randomized to either oral estrogen (conjugated equine estrogens 0.625 mg/day) plus progestin (medroxyprogesterone acetate 2.5 mg/day) or placebo. There was a wash-out interval of 3 months before randomization.

Our GPRD trials included women who either initiated oral therapy with estrogens plus progesterone or were nonusers of hormone therapy in month m. As an additional eligibility criterion, in each trial m, women were required to have been nonusers of any form of hormone therapy during the year before baseline (wash-out interval). (We choose a year rather than 3 months to hopefully obtain a better match with the WHI on the distribution of "time since last hormone therapy.") We refer to women eligible for trial m who did (did not) initiate hormone therapy in month m as "initiators" (noninitiators) in trial m.

2.4 *Ascertainment of CHD Endpoints and Confounding Variables*

As in the original GPRD analysis (Varas-Lorenzo et al., 2000), we defined the CHD endpoint in study m as the time of nonfatal myocardial infarction or fatal coronary disease between baseline (as defined above) and end of follow-up. The follow-up in the original GPRD study ended in December 1995. Our reanalyses extend follow-up to December 2001. In the original study, over 90% of CHD endpoints ascertained after review of computer records were confirmed by reviewing the patients' paper medical records and using standardized diagnostic criteria.

In each trial m, we obtained at baseline (i.e., just prior to month m) data on the following potential confounders: age, calendar month, family history of CHD, high cholesterol, high blood pressure, diabetes, body mass index, smoking, alcohol intake, aspirin use, nonsteroidal anti-inflammatory drug use, and previous use of hormone therapy. Data on additional potential "lifestyle" confounders were unavailable.

3. Analytic Approach for the GPRD Trials

As discussed further below, our conceptualization of an observational study with a time-varying treatment as a sequence of trials can be viewed as a special case of g-estimation of nested structural models (Robins, 1989).

3.1 *ITT Effect of Treatment*

In each GPRD trial, we compared the CHD hazard rate of initiators and noninitiators, regardless of whether these women subsequently stopped or initiated therapy. Thus our approach is the observational equivalent of the ITT principle that guided the main analysis of the WHI trial. To the women eligible for each GPRD trial m, we fit the Cox proportional hazards model

$$\lambda_T\big[t \,|\, G(m)=1, A(m), \bar{L}(m)\big] = \lambda_0\big[t\big]\, \exp\big[\alpha A(m) + \eta \bar{L}(m)\big], \quad (1)$$

where m indexes the trial (months from January 1991), T is the time from baseline of trial m to CHD, $G(m)$ is an indicator for eligibility for trial m (1: yes, 0: no), $A(m)$ is hormone therapy initiation at m (1: yes, 0: no), $\bar{L}(m)$ is a vector representing covariate history through baseline m, $\lambda_T[t \,|\, G(m)=1, \bar{L}(m), A(m)]$ is the conditional hazard of CHD at time t, $\lambda_0[t]$ is the baseline hazard at t, and $\exp[\alpha]$ is the conditional ITT hazard ratio for hormone therapy initiation versus noninitiation at baseline m. We modeled $\bar{L}(m)$ by including the potential confounders described in the previous section as covariates. All covariates were categorical except age, alcohol intake, and calendar month. The age effect was modeled as cubic splines with 3 knots and with product terms of the age coefficients with diabetes and hypertension. To increase precision, we pooled all 132 GPRD trials in a single analysis. Because many women participate in more than one trial, we used the robust variance to account for within-person correlation. In addition to our main analyses, we conducted subgroup analyses by age (<60, ≥ 60 years) at baseline and investigated how the rate ratio exp (α) was modified by the month m of the trial and by time since initiation of therapy.

Under the assumption of no unmeasured confounders, our Cox model estimates the conditional ITT hazard ratio $\exp[\alpha]$ within levels of $\bar{L}(m)$, that is, the (conditional) hazard had everybody initiated treatment divided by the hazard had nobody initiated treatment in each GPRD trial. Note that when this analytic approach is applied to a closed cohort in which noneligible women never become eligible at later times, each trial is nested in the prior trial (Hernán et al., 2005) and we refer to the Cox model (1) pooled over all trials as a nested Cox model.

3.2 *Effect of Continuous Treatment*

The magnitude of the ITT hazard ratio in a study depends not only on the effect of hormone therapy but also on the degree of "compliance." (In our GPRD trials, we defined the time to noncompliance in trial m as the difference between m and the month of first deviation from baseline treatment, i.e., discontinuation of hormone therapy for initiators, and initiation of hormone therapy for noninitiators.) The WHI and the GPRD differ markedly in their "time to noncompliance" distributions (see Section 5 below), which could cause their ITT hazard ratios to differ substantially. To disaggregate the effect of noncompliance from the effect of hormone therapy, we attempted to estimate for the GPRD trials the "continuous treatment hazard ratio" that would be observed under full compliance, that is, the hazard ratio comparing continuous treatment in all initiators versus no treatment in all noninitiators.

To do so, separately in each trial m, we censored women when they discontinued their baseline treatment. Because this censoring is potentially informative (i.e., noncompliance is nonrandom) and may lead to selection bias (Hernán, Hernández-Díaz, and Robins, 2004), a women i at risk (and thus uncensored) in month $k > m$ was upweighted by the inverse of her estimated probability of remaining uncensored from month m through month k. Specifically, for each trial m we fit logistic models

$$\begin{aligned}\text{logit Pr}\big[A(j)=a \,|\, G(m)=1,& \\ A(j-1)=a, A(m)=a, \bar{L}(j), T>j\big]& \\ = \theta_{a0} + \theta'_{a1}\bar{L}(j) \quad \text{for} \quad j>m,& \quad (2)\end{aligned}$$

for continuing compliance separately for initiators ($a=1$) and noninitiators ($a=0$). The estimated probability of continuing the baseline treatment through month $k > m$ for subject i is the product $\Pi_{j=m+1}^{k}\hat{P}_{mi}(j)$ where $\hat{P}_{mi}(j)$ is the predicted value

$$\begin{aligned}\hat{P}_{mi}(j) = G_i(m)\widehat{\Pr}\big[A(j)=a \,|\, G(m)=1, A(j-1)=a& \\ A(m)=a, \bar{L}_i(j), T>j\big]_{|a=A_i(m)},&\end{aligned}$$

from the logistic models. We then estimated the rate ratio $\exp[\alpha]$ by refitting Cox model (1) after censoring them at the time of discontinuation of baseline treatment and weighting their contributions to the partial likelihood at time k by the inverse probability weights (IPW) $\hat{W}_{m,i}(k) = [\prod_{j=m+1}^{k}\hat{P}_{mi}(j)]^{-1}$. Again, to increase precision we pooled all 132 GPRD trials in a single analysis. The assumptions required for the limit of $\exp[\hat{\alpha}]$ to be the "continuous treatment hazard ratio" are discussed in Section 5. To examine whether censoring due to noncompliance was "informative," we repeated the above analysis without weights (i.e., we set all the $\hat{W}_{m,i}(k)$ to 1).

For comparison purposes, we will also fit a standard time-varying Cox model

$$\lambda_{T'}\big[t \,|\, G(0)=1, A(t), \bar{L}(t)\big] = \lambda_0[t]\exp\left[\beta_c A_c\,(t) + \beta_p A_p\,(t) + \gamma' L(t)\right], \quad (3)$$

where T' is the time from the first eligible trial (i.e., month) to CHD, A_c is an indicator for being currently on treatment, A_p is an indicator for being a past user at t (past treatment), and $L(t)$ are the updated covariate values at t. The hazard ratios $\exp[\beta_c]$ and $\exp[\beta_p]$ compare the CHD incidence in

Table 1
Number of participants, hormone therapy initiators, and CHD events in each GPRD trial (for illustration purposes, only trials 25–50 are shown)

Trial	Month	Participants	CHD events	Initiators	CHD events in initiators
25	January 1993	68,026	1134	218	1
26	February 1993	67,774	1112	193	1
27	March 1993	67,669	1085	239	1
28	April 1993	67,338	1060	201	1
29	May 1993	66,972	1030	200	1
30	June 1993	66,893	1009	170	1
31	July 1993	66,720	985	168	0
32	August 1993	66,655	966	192	0
33	September 1993	66,354	947	134	1
34	October 1993	66,301	928	132	0
35	November 1993	66,165	908	155	1
36	December 1993	65,983	884	98	0
37	January 1994	69,729	871	149	2
38	February 1994	69,592	858	185	2
39	March 1994	69,262	833	196	3
40	April 1994	69,019	813	168	0
41	May 1994	68,919	801	141	0
42	June 1994	68,442	785	146	1
43	July 1994	68,245	751	135	0
44	August 1994	68,053	736	158	0
45	September 1994	67,769	718	137	2
46	October 1994	67,681	689	135	1
47	November 1994	67,413	661	145	1
48	December 1994	67,151	648	97	0
49	January 1995	69,901	626	178	1
50	February 1995	69,500	618	146	1

current and past users at t, respectively, with that of never users within levels of the updated covariates $L(t)$. We investigated how the rate ratio at t was modified by the duration $D(t)$ since the last reinitiation of hormone therapy (following a period of at least a year of nonuse) at an eligible month by adding, for example, $\beta_1 A_c(t)\ I(5 > D(t) > 2) + \beta_2 A_c(t)\ I(D(t) > 5) + \beta_3 A_c(t)\ N(t)$ to the model, where $N(t)$ is one if a subject never initiated therapy at an eligible month and zero otherwise.

4. Results from the GPRD Trials

Our analyses included 99,072 women who met the eligibility criteria for at least one GPRD trial. Of these women, 1889 had a CHD event and 606 died during the follow-up.

On average, each woman participated in 60.5 trials (standard deviation [SD]: 35.3, median: 59.0) and thus our analyses include 5,997,824 (nondistinct) women, 10,566 initiators, 64,583 CHD endpoints, and 20,815 deaths when all trials are pooled. The records of 16% of the initiators and 9% of the noninitiators indicated use of hormone therapy more than 1 year before baseline. Only 64 CHD endpoints occurred among initiators, thus limiting the precision of our analysis. As an example, Table 1 shows the number of participants, initiators, and CHD events in trials 25–50. The mean duration of follow-up across all trials was 4.1 years (SD: 2.6, median: 3.8 years), and the mean age at baseline was 54.6 years (SD: 4.5, median: 53.0) for the initiators and 62.0 years (SD: 6.7, median: 62.0) for the noninitiators.

4.1 *Estimates of the ITT Effect*

The estimated ITT hazard ratio (95% CI) of CHD for hormone therapy initiation versus no initiation from model (1) was 0.92 (0.73, 1.17). When an interaction term between treatment initiation at baseline $A(m)$ and the month m that the trial began was added, the term's estimated coefficient (95% CI) was 0.005 (−0.059, 0.158), indicating little evidence of trial time–treatment interaction. The estimated ITT hazard ratios (95% CI) were 0.97 (0.74, 1.27) for women younger than 60 years and 0.73 (0.44, 1.22) for women 60 years and older at baseline. Table 2 shows the estimates when we restricted the analysis to various periods of follow-up. A further breakdown shows hazard ratios of 0.82 (0.55, 1.21) in years 2–5, and 0.69 (0.38, 1.25) in years 5–10.

We also estimated the ITT effect of hormone therapy on mortality after replacing "time to CHD" by "time to death" in model (1). The estimated ITT hazard ratio (95% CI) of death for hormone therapy initiation versus no initiation was 0.89 (0.55, 1.46). When we restricted the duration of the GPRD trials, the respective estimated ITT hazard ratios (95% CI) were 1.27 (0.66, 2.43) for 2 years, 1.09 (0.67, 1.77) for 5 years, and 0.90 (0.75, 1.20) for 8 years.

4.2 *Estimates of the Continuous Treatment Effect and Standard Covariate-Updated Analyses*

The proportion of noninitiators who initiated therapy (Figure 1) and of initiators who discontinued therapy (Figure 2) increased over the follow-up period. By 6 years

Table 2
CHD hazard ratios and 95% confidence intervals for hormone therapy use in the GPRD trials

Years of follow-up	Initiators versus noninitiators Model (1) ITT	Continuous versus never users Model (1) IPW	Unweighted	Current versus never users Model (3) Updated covariates
0–2	1.20 (0.84, 1.72)	1.33 (0.79, 2.22)	1.32 (0.82, 2.13)	1.02 (0.63, 1.65)
0–5	0.99 (0.76, 1.28)	0.83 (0.52, 1.32)	0.98 (0.65, 1.49)	0.80 (0.52, 1.21)
0–8	0.95 (0.75, 1.20)	0.95 (0.60, 1.51)	0.98 (0.67, 1.43)	0.88 (0.61, 1.28)
All	0.92 (0.73, 1.17)	0.87 (0.55, 1.39)	0.97 (0.66, 1.42)	0.87 (0.60, 1.27)

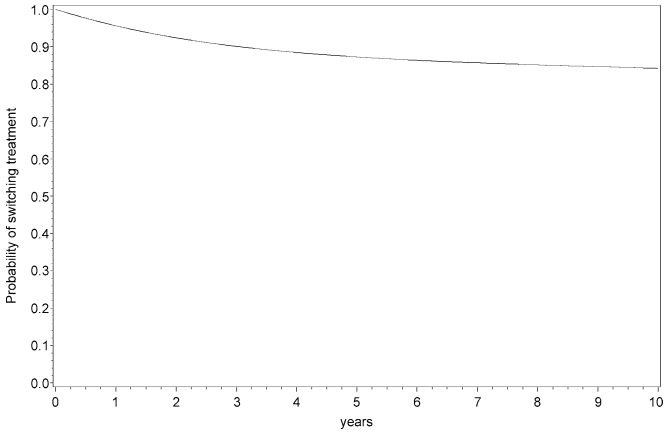

Figure 1. Probability of initiating hormone therapy among noninitiators.

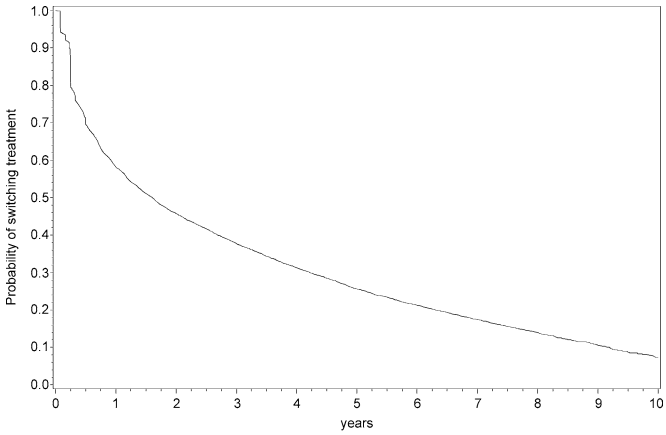

Figure 2. Probability of discontinuing hormone therapy among initiators.

of follow-up, the proportion of noncompliance was 13% in noninitiators and 79% in initiators. In the latter group, the steepest drop in hormone therapy use occurred during the first year after baseline (Figure 2). The high discontinuation rate found in the GPRD reflects that of the general British population (Bromley, de Vries, and Farmer, 2004).

Using IPW to adjust for informative censoring, the estimated hazard ratio of CHD for continuous hormone therapy versus no therapy using weighted model (1) was 0.87 (0.55, 1.39). When we did not weight, we obtained an estimated hazard ratio of 0.97 (0.66, 1.42). Table 2 shows the weighted and unweighted estimates when we restricted the analysis to various periods of follow-up.

The standard updated-covariate analysis gave an estimated hazard ratio 0.87 (0.60, 1.27) for current therapy was never exposed since the first eligible visit. The last column of Table 2 shows the corresponding covariate-updated estimates as a function of duration of treatment (since the last eligible period). A further breakdown shows hazard ratios of 0.48 (0.21, 1.06) in years 2–5, and 1.34 (0.60, 3.01) in years 5–10.

5. Discussion and Comparison with Prentice et al.

Our ITT analysis of our GPRD trials suggest that initiation of estrogen plus progestin does not have a substantial impact on the risk of CHD although, when compared with noninitiators, the CHD incidence of initiators was 20% greater during the 2-year period after initiation and 5% lower when averaged over the 8-year period after initiation. Neither estimate approached statistical significance.

We did not find significant risk differences by age, but power was limited because few younger women had a CHD endpoint and few older women initiated therapy. We could not stratify the analysis by time since menopause because time of menopause is not systematically recorded in the GPRD. When we further restricted eligibility in trial m by requiring no prior recorded hormone use (rather than a year wash-out period), 91% of the previously eligible women remained eligible and the ITT estimates showed little change (data not shown).

The ITT estimates from the GPRD trials are closer to the null than those of the WHI trial (WHI overall hazard ratio: 1.24, 95% CI: 0.97, 1.60) (Manson et al., 2003). This attenuation may be a consequence of the presence of unmeasured confounding for treatment initiation in the GPRD, a higher proportion of noncompliance in the GPRD trials, random variability in both studies, or a combination of these factors. The GPRD-WHI ITT differences cannot be explained by any uncertainty in time of therapy initiation or by confounding by risk factors whose distribution differed in women who continued versus discontinued therapy.

Our approach provides unbiased estimates of the ITT effect only under the assumption of no unmeasured confounders for treatment initiation. Although this assumption cannot be directly tested in observational studies, comparison between the adjusted and the unadjusted estimates can be useful in assessing the hypothesis that substantial confounding by unmeasured factors remains. When we repeated our ITT analysis without adjustment for baseline covariates (except age and

calendar month), the estimated hazard ratio was 0.85 (95% CI: 0.67, 1.08), which is only moderately less than the fully adjusted estimate 0.92. Were sampling variability absent, it would then follow that the magnitude of confounding due to unmeasured variables would have to exceed the confounding due to measured variables to explain the full GPRD-WHI discrepancy. Given the breadth of the measured variables, we believe this hypothesis seems unlikely, although a downward bias of perhaps 0.1 or 0.2 in our hazard ratio estimate is still plausible, especially in light of the large sampling variability. Indeed large sampling variability is a major problem. For example, the overall ITT hazard ratios from the GPRD and the WHI trials were estimated with similarly low precision (width of the 95% CIs on the log scale: about 0.46 in WHI and 0.47 in GPRD) with point estimates close to the null. This relatively low precision precludes drawing strong conclusions from either study and produces overlapping confidence intervals for the GPRD and the WHI estimates and a nonsignificant estimated difference in ITT effects. For the all-cause mortality hazard our GPRD estimates were quite similar to the WHI estimate of 0.98 (95% CI: 0.70, 1.37).

Both the WHI and our primary GPRD analysis estimated the ITT effect of hormone therapy initiation rather than the effect of continuing hormone therapy. Because the rate of noncompliance differed between the GPRD and the WHI trials (Writing Group for the WHI Investigators, 2002): 42% (WHI) versus 79% (GPRD) in initiators, and 11% (WHI) versus 13% (GPRD) in noninitiators at 6 years of follow-up, our GPRD estimates may not be directly comparable with the WHI estimates. To eliminate the effect of noncompliance, we attempted to estimate the "continuous treatment or full compliance" hazard ratio (i.e., the ITT effect in the absence of noncompliance) in the GPRD trials by IPW. As discussed by Robins and Finkelstein (2000) and Robins (1998), one should not regard as noncompliant women whose deviation from their assigned therapy was for (not easily palliated) adverse medical reasons. Prentice et al. make a similar point. However, in the GPRD study, this option was not available to us, as data on why a woman stopped hormone therapy was not routinely collected. Robins and Finkelstein (2000) showed that the IPW estimates are consistent for the "continuous treatment" hazard ratio if (i) women who initiated and did not initiate hormone therapy in each trial m were comparable conditionally on $\bar{L}(m)$ (no unmeasured confounding for treatment), (ii) women who discontinued and did not discontinue their baseline treatment in each month k were comparable conditionally on $\bar{L}(k)$ (no unmeasured confounding for censoring), and (iii) model misspecification is absent. Our IPW methodology to correct for informative censoring is a special case of the much more general methodology of IPW estimation of marginal structural models. In the GPRD, the overall IPW hazard ratio estimate of 0.87 was close to the overall ITT estimate of 0.92.

Further, by comparing the weighted and unweighted estimates of the continuous therapy effect in Table 2, we can see that although censoring by noncompliance may have been moderately informative, the observed differences are not statistically significant. It would be interesting to conduct IPW analyses of censoring by noncompliance in the WHI trial as well.

Although we did not do so here, in the presence of unmeasured confounding for treatment continuation (i.e., continued compliance), IPW estimators can be used to conduct a sensitivity analysis as follows. Suppose, for the moment, the amount of unmeasured confounding were known, in the sense that we could choose a parameter ω and a function $q(j, m, \bar{L}(j), T_{\bar{a}})$ such that their product $\omega q(j, m, a, \bar{L}(j), T_{\bar{a}})$ correctly quantifies the degree of dependence on the log-odds scale between the probability of treatment continuation and the counterfactual survival time $T_{\bar{a}}$ under treatment history $\bar{a}$ through the model

$$\begin{aligned}\text{logit Pr}\big[A(j) = a \,|\, G(m) = 1, A(j-1) = a,\\ A(m) = a, \bar{L}(j), T > j, T_{\bar{a}}\big]\\ = \theta_{a0} + \theta'_{a1}\bar{L}(j) + \omega q(j, m, a, \bar{L}(j), T_{\bar{a}}) \quad \text{for} \quad j > m. \quad (4)\end{aligned}$$

This logistic model reduces to model (2) if there were no unmeasured confounding for continued compliance (i.e., $\omega q(j, m, a, \bar{L}(j), T_{\bar{a}}) = 0$). Because the degree of unmeasured confounding is actually unknown, we suggest a sensitivity analysis in which one plots estimates and confidence intervals for the "continuous treatment" hazard ratio as a function of ω and $q(j, m, a, \bar{L}(j), T_{\bar{a}})$, where ω and $q(j, m, a, \bar{L}(j), T_{\bar{a}})$ are allowed to vary over a plausible range of values and functional forms (Scharfstein et al., 2001; Robins, 2002).

Prentice and collaborators also consider estimating the full compliance hazard ratio in the WHI randomized trial by censoring subjects at the time of noncompliance, but do not use data on evolving postrandomization covariates $\bar{L}(j)$ to reweight subjects. Prentice and collaborators conjecture that any bias due to this failure to adjust for $\bar{L}(k)$ is likely small. The rather modest differences in weighted and the unweighted estimates in Table 2 serve as an empirical test and partial confirmation of this conjecture in the GPRD. However, in observational studies of the effect of drug therapy on time to AIDS or death in HIV-infected subjects, the magnitude of confounding by time-varying covariates (e.g., CD4 cell count) is much larger than for the effect of hormone therapy on CHD in the GPRD study. In these studies we have repeatedly shown that standard analytic strategies fail; only "causal inference" methods (either IPW estimation of marginal structural models or *g*-estimation of nested structural models) successfully reproduce the results of randomized trials (Cole et al., 2003; Hernán et al., 2005; Sterne et al., 2005). Because small bias cannot be assured a priori, we believe an analyst should routinely correct (separately in each arm) for selection bias attributable to the measured factors $\bar{L}(k)$ by using IPW and should, perhaps, also consider using IPW to investigate the sensitivity of one's inferences to confounding by unmeasured factors.

Prentice et al. mention the existence of methods for analyzing double blind randomized trial suffering from noncompliance that both (i), like an as treated analysis, provide estimates of the treatment effect under full compliance and yet (ii), like an ITT analysis, protect the α-level under the null hypothesis of no treatment effect (without imposing any assumptions concerning either the existence or magnitude of unmeasured confounding for treatment continuation). Specifically, Prentice et al. reference methodologies proposed

by Cuzick, Edwards, and Segnan (1997) and Frangakis and Rubin (1999). However, these methodologies only apply if compliance is of the "all or none" type, and censoring by end of follow-up is independent conditional on complier type. But in the WHI compliance is complex and time varying with women repeatedly stopping and starting their assigned therapy. Further, although less likely, censoring by end of follow-up may be dependent if secular changes in baseline mortality risk have occurred over the trial accrual period. In this setting, as far as we are aware, *g*-estimation of nested structural models (usually referred to as SNFTMs) is the only general methodology available for the analysis of failure time data that satisfies both (i) and (ii) (Mark and Robins, 1993). Of course, adequate data on actual treatment $A(t)$ must be available for analysis. The Appendix provides further detail.

We could have also used doubly robust *g*-estimation of an SNFTM rather than our IPW methodology to estimate the effect of continuous hormone therapy on CHD in the GPRD study. Doubly robust *g*-estimation provides consistent estimation of the effect of continuous hormone therapy if there is no unmeasured confounding for treatment initiation, the SNFTM is correct, and one has correctly specified either (but not necessarily both) a model for the conditional probability that an eligible subject (i.e., $G(m) = 1$) initiates treatment in trial m given $\bar{L}(m)$ or a model for the counterfactual regressions $E[T_{m,0} \mid \bar{L}(m), G(m) = 1, T > m]$ where $T_{m,0}$ is a subject's possibly counterfactual time to CHD had the subject received her observed treatment $\bar{A}(m-1)$ up till month $m-1$ and no treatment from m (these *g*-estimators are referred to as doubly robust because of this latter property). The requirement for correct specification of the SNFTM in *g*-estimation substitutes for the requirement for correct specification of model (4) in IPW estimation.

Furthermore, we could have used doubly robust *g*-estimation of an SNFTM to estimate the ITT effect of treatment in our GPRD trials but on a multiplicative survival scale rather than on a hazard ratio scale. In this setting the simplest SNFTM is a nested AFT model (defined in the Appendix). A nested AFT model (and more generally any ITT SNFTM) has certain theoretical advantages compared with nested hazard ratio models such as the nested Cox model that we used. First, as remarked by Prentice et al., and in contrast with nested AFT models, if the treatment and control ITT hazards cross at some time t, the values of the parameters of even a correctly specified ITT hazard ratio model do not determine when (or even whether) the survival curves also cross, unless combined with an estimate of the baseline survivor function. (Only when the survival curves cross can one logically conclude that treatment benefits some subjects and harms others.) Second, standard hazard ratio models do not admit doubly robust estimators, although this shortcoming in robustness can be alleviated by using marginal structural hazard ratio models.

Reading from Table 2, we see that there is no qualitative difference between IPW results and the results from the standard updated-covariate analysis, especially in view of the substantial sampling variability. Both analyses suggest a possible hazard ratio of less than 1 when the duration of therapy is from 2 to 5 years. However in light of the large sampling variability and multiple comparison considerations, no definitive conclusions are possible. In contrast with the qualitative agreement in the GPRD, in studies of the effect of highly active antiretroviral therapy (HAART) on (i) time to AIDS or death and (ii) on evolution of CD4 count in HIV-infected subjects, IPW succeeded but standard updated-covariate analyses failed to reproduce results found in randomized clinical trials. The problem with the standard updated-covariate analysis is that it adjusts for covariates affected by earlier treatment, which can result in bias (Hernán et al., 2004).

As mentioned in the introduction, the original standard updated-covariate analyses of the GPRD reported a statistically significant hazard ratio of 0.72 (0.59, 0.89) for current versus never exposed. However, the original 1995 GPRD analyses differed from ours in that (i) all hormone users (including estrogen only users) were compared to never users, (ii) a subject was defined as "currently" exposed at t if exposed any time in the 6 months before t (regardless of past use history), and (iii) the maximum duration of follow-up was 5 years rather than 10 years. When we repeated our analyses using definition (ii) of current exposure, effect estimates were little changed (data not shown). As discussed above, our analyses suggest (but do not prove) that the hazard ratio is modified by duration of exposure and thus presumably by duration of follow-up when current exposure is coded simply as 1 or 0. Thus the difference between our results and those of the original GPRD analyses are presumably due to (i) and perhaps to (iii).

Finally, five remaining differences may affect the GPRD-WHI comparability. First, individuals in the GPRD trials were not blinded as to whether they did or did not receive hormone therapy. If awareness of exposure status modified the behavior of either the women or their physicians in ways that affected the risk of a CHD diagnosis, then the GPRD estimates would reflect the joint effect of hormone therapy and these behavioral modifications. WHI participants were initially blinded to treatment regime, although some of them may have become aware of it later on, and in fact differential unblinding of hormone users has been suggested as a potential source of bias in the WHI (Garbe and Suissa, 2004). Second, women with conditions inconsistent with adherence (e.g., menopausal symptoms) were excluded in the WHI but not in our GPRD analysis. The GPRD and WHI results might differ if, as the WHI results suggest (Manson et al., 2003), hormone therapy is less harmful, or possibly beneficial, in the presence of menopausal symptoms. Third, women who initiated hormone therapy in the GPRD were, on average, 8.6 years younger than initiators in the WHI. Fourth, the particular drugs used for postmenopausal hormonal therapy in the WHI and in the GPRD are different. Last, there is no guarantee that the GPRD and WHI noncompliers were comparable. For example, many of the GPRD "noncompliers" stopped hormone therapy simply because their physician prescribed the drug only for a brief period to combat menopausal symptoms. This last concern could be partly alleviated by comparing the effects of continued hormone therapy in both the GPRD and the WHI using either IPW or *g*-estimation methodology.

In conclusion, we have described an analytic approach for observational studies that mimics that commonly used for randomized trials and that allows more direct comparisons between the results of observational and randomized studies. Under our approach no clear beneficial effect or adverse effect of combined hormone therapy is apparent in the GPRD, but we had little power to discover small to moderate effects. The difference between the overall WHI ITT estimate of 1.24 and our GPRD ITT estimate of 0.92 is consistent with random variability, although additional systematic sources of small to moderate bias cannot be excluded in the GPRD. Unfortunately, because of the large sampling variability in both the WHI trial and the GPRD study, our results shed little light on the question of whether an (even correctly analyzed) observational study of a "lifestyle exposure" can reliably discriminate among causal relative risks close to 1. Prentice et al. show that when the hazard ratio is allowed to vary with duration of therapy, the WHI randomized trial and the WHI observational study provide similar hazard ratio estimates. But these authors also had little power to distinguish this similarity hypothesis from the hypothesis of a moderate systematic difference between the hazard ratios, which raises the following counterfactual questions that we hope the authors might respond to in their rejoinder. Had the WHI randomized trial been cancelled and the only data been that from the WHI observational study, would Prentice et al. have analyzed the data in the same way and reached the same conclusions as in their actual paper? Further, what is their best explanation of the discrepancy between the results of their WHI observational analysis and the results of the other observational studies that found a clear benefit of hormone therapy on CHD? How certain are they that this explanation is correct? We ask because, in our analyses of the GPRD and the NHS, we have often been unable to find clear and convincing explanations for the variation observed in our effect estimates with elaboration of the analytic model in different directions.

References

Bromley, S. E., de Vries, C. S., and Farmer, R. D. T. (2004). Utilisation of hormone replacement therapy in the United Kingdom. A descriptive study using the general practice research database. *British Journal of Obstetrics and Gynaecology* **111,** 369–376.

Cole, S. R., Hernán, M. A., Robins, J. M., et al. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology* **158,** 687–694.

Cuzick, J., Edwards, R., and Segnan, N. (1997). Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in Medicine* **16,** 1017–1029.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment compliance and subsequent missing outcomes. *Biometrika* **86,** 365–379.

Garbe, E. and Suissa, S. (2004). Hormone replacement therapy and acute coronary outcomes: Methodological issues between randomized and observational studies. *Human Reproduction* **19,** 8–13.

García Rodríguez, L. A. and Pérez Gutthann, S. (1998). Use of the UK General Practice Research Database for pharmacoepidemiology. *British Journal of Clinical Pharmacology* **45,** 419–425.

Grodstein, F., Stampfer, M. J., Manson, J. E., Colditz, G. A., Willett, W. C., Rosner, B., Speizer, F. E., and Hennekens, C. H. (1996). Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. *New England Journal of Medicine* **335,** 453–461. (Erratum in *New England Journal of Medicine* 1996, **335,** 1406.)

Grodstein, F., Manson, J. E., Colditz, G. A., Willett, W. C., Speizer, F. E., and Stampfer, M. J. (2000). A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. *Annals of Internal Medicine* **133,** 933–941.

Grodstein, F., Manson, J. E., and Stampfer, M. J. (2001). Postmenopausal hormone use and secondary prevention of coronary events in the Nurses' Health Study. A prospective, observational study. *Annals of Internal Medicine* **135,** 1–8.

Grodstein, F., Clarkson, T. B., and Manson, J. E. (2003). Understanding the divergent data on postmenopausal hormone therapy. *New England Journal of Medicine* **348,** 645–650.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology* **15,** 615–625.

Hernán, M. A., Cole, S. R., Margolick, J. B., Cohen, M. H., and Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety* **14,** 477–491.

Jick, S. S., Kaye, J. A., Vasilakis-Scaramozza, C., García Rodríguez, L. A., Ruigómez, A., Meier, C. R., Schlienger, R. G., Black, C., and Jick, H. (2003). Validity of the General Practice Research Database. *Pharmacotherapy* **23,** 686–689.

Manson, J. E., Hsia, J., Johnson, K. C., et al. and the Women's Health Initiative Investigators. (2003). Estrogen plus progestin and the risk of coronary heart disease. *New England Journal of Medicine* **349,** 523–534.

Mark, S. D. and Robins, J. M. (1993). A method for the analysis of randomized trials with compliance information: An application to the multiple risk factor intervention trial. *Controlled Clinical Trials* **14,** 79–97.

Mendelsohn, M. E. and Karas, R. H. (2005). Molecular and cellular basis of cardiovascular gender differences. *Science* **308,** 1583–1587.

Robins, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Services Research Methodology: A Focus on AIDS*, L. Sechrest, H. Freeman, and A. Mulley (eds), 113–159. Washington, DC: U.S. Public Health Service, National Center for Health Services Research.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine* **17,** 269–302.

Robins, J. M. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies" by Paul R. Rosenbaum. *Statistical Science* **17,** 286–327.

Robins, J. M. and Finkelstein, D. (2000). Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56,** 779–788.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients. *Epidemiology* **3,** 319–336. (Erratum in *Epidemiology* 1993, **4,** 189.)

Scharfstein, D. O., Robins, J. M., Eddings, W., and Rotnitzky, A. (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints. *Biometrics* **57,** 404–413.

Stampfer, M. J., Colditz, G. A., Willett, W. C., Manson, J. E., Rosner, B., Speizer, F. E., and Hennekens, C. H. (1991). Postmenopausal estrogen therapy and cardiovascular disease. Ten-year follow-up from the Nurses' Health Study. *New England Journal of Medicine* **325,** 756–762.

Sterne, J. A. C., Hernán, M. A., Ledergerber, B., Tilling, K., Weber, R., Robins, J. M., and Egger, M., the Swiss HIV Cohort Study. (2005). Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: The Swiss HIV Cohort Study. *Lancet* **366,** 378–384.

Varas-Lorenzo, C., García-Rodríguez, L. A., Pérez-Gutthann, S., and Duque-Oliart, A. (2000). Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. *Circulation* **101,** 2572–2578.

Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. *Controlled Clinical Trials* **19,** 61–109.

Writing Group for the Women's Health Initiative Investigators. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* **288,** 321–333.

Appendix

G-Estimation of Nested Structural Models for Survival Analysis

The simplest structural nested failure time model (SNFTM) implies that for some unknown value ψ^* of ψ, the observable random variable $H_m(\psi) = h_m(T, \bar{A}(T), \psi) = \int_m^T \exp(\psi A(t))\, dt$ has a conditional distribution given $(\bar{L}(m), \bar{A}(m), T > m)$ equal to that of $T_{m,0}$, where $T_{m,0}$ is defined in the main text. This model is related to the time-dependent accelerated failure time model. It implies that for each m, $(T_{m,1} - m)$ has the same distribution as $\exp(-\psi^*)$ $(T_{m,0} - m)$ and where $T_{m,1}$ is a subject's possibly counterfactual time to CHD had the subject received her observed treatment $\bar{A}(m-1)$ up to month m and continuous treatment from m onward. In particular, continuous treatment from $m = 0$ scales the survival distribution by a factor $\exp(-\psi^*)$ compared to no treatment. The parameter ψ^* is estimated with doubly robust *g*-estimation. A general SNFTM posits $H_m(\psi) = h_m(T, \bar{A}(T), \bar{L}(T), \psi)$ to be a known function of $(T, \bar{A}(T), \bar{L}(T), \psi)$ increasing in T and satisfying $H_m(\psi) = T - m$ if $\psi = 0$ or $A(u) = 0$, $m \le u < \infty$. Robins et al. (1992) extends *g*-estimation to allow for right censoring both by administrative end of follow-up and by competing risks. Owing to double robustness and to the fact that structural nested failure time models are guaranteed correct (with true $\psi^* = 0$) whenever a hormone effect on CHD is absent, *g*-estimation can be used to construct robust tests of the null hypothesis of no effect of hormone therapy on CHD, whenever there is no unmeasured confounding for treatment initiation.

To estimate the ITT effect of therapy, we simply redefine $H_m(\psi) = \int_m^T \exp(\psi A(m))\, dt = A(m) \exp(\psi)$, then $\exp(-\psi)$ now has the meaning of the ITT effect of treatment, assumed to be the same for each trial m. We refer to this model as a time-independent nested AFT model for the ITT effect. A general *ITT* SNFTM has $H_m(\psi) = h_m(T, \bar{A}(m), \bar{L}(m), \psi)$ with $h_m(T, \bar{A}(m), \bar{L}(m), \psi) = T - m$ if $\psi = 0$ or $A(m) = 0$.

Duncan C. Thomas

University of Southern California
Preventative Medicine (Division of Biostatistics)
Los Angeles, California, Los Angeles, CA
email: *dthomas@usc.edu*

In a typically masterful performance, Prentice, Pettinger, and Anderson have beautifully summarized a broad range of statistical issues raised by one of the most important longitudinal studies of our day, combining both randomized trial and observational epidemiology components. As other commentators will address many of the clinical trial and observational epidemiology issues, I will confine my remarks to the genetic issues raised in Section 2.2. In particular, I will focus on the discussion of germline variation, although many of the problems associated with very high density data arising in that context also apply to the proteomic data. This is frequently referred to as the "$p \gg n$ problem," meaning many more variables than observations.

To begin with, the focus of Prentice et al.'s discussion of germline variation is on detecting the main effects of genetic variants on disease risk. Many such "genome-wide association scans" (GWASs)—first seriously proposed nearly a decade ago by Risch and Merikangas (1996)—have recently been proposed and some are already underway (see review of several such initiatives in Thomas, Haile, and Duggan, 2005). Indeed, the first reports of such scans have started to appear (Ozaki et al., 2002; Klein et al., 2005; Maraganore et al., 2005). Before discussing some of the methodological issues involved in GWASs, it's worth noting that much of the interest in the pharmacogenomics world centers on genetic *modifiers* of the response to drug treatments (Need, Motulsky, and Goldstein,

2005), or in the case of the Women's Health Initiative (WHI), on genetic modifiers of chemopreventive agents. A timely reminder of the importance of such research is the approval by the U.S. FDA on June 16, 2005 of the drug BiDil (NitroMed, `http://www.nitromed.com/index.asp`) for treatment of congestive heart failure only in African-Americans. As noted by Francis Collins and other critics of this decision, it is highly unlikely that race or skin color per se is the relevant factor modifying the effectiveness in the treatment (if indeed there really is a racial difference), but rather some as-yet-undiscovered genetic variant that is more prevalent among African-Americans (Kahn, 2005). A search for such a modifier gene may prove to be a more daunting challenge than for a main effect, but as noted by the FDA panel's chairman, Dr Steven E. Nissen, "It is very unusual; it is precedent-setting, but it is the case that we are moving forward to genome-based medicine. It's going to happen."

The generally greater difficulty in detecting interactions than main effects in observational studies (Smith and Day, 1984) is somewhat offset in the pharmacogenetics field, however, by the ability to randomize one of the interacting factors, in this case the exposure (drug). Although this does not alter the sample size requirements (other than ensuring a suitable balance of exposed and unexposed), it does provide a stronger basis for causal inference. In a similar vein, prospects for causal inference could be enhanced by exploiting the concept of "Mendelian randomization" (Davey Smith and Ebrahim, 2003; Little and Khoury, 2003; Thomas and Conti, 2004), meaning that genes are "assigned" to individuals randomly conditional on parental genotypes, so that issues of "reverse causation" (disease influencing an intermediate phenotype under study) and confounding can largely be eliminated as alternative explanations. To fully exploit this advantage, family-based designs that use such transmission information could be used, such as the transmission-disequilibrium test (TDT) (Spielman, McGinnis, and Ewens, 1993) or discordant sibship case–control design (Kraft and Thomas, 2004). While the latter design can be impractical in a randomized trial context, a TDT analysis could take the form of a comparison of transmitted genotypes between affected cases on differing treatment arms. Although this would require the availability of DNA from both parents of all cases, it would not require unaffected controls, thereby substantially reducing genotyping costs and ensuring freedom from population stratification bias (Thomas and Witte, 2002; Wacholder, Rothman, and Caporaso, 2002). The ideas of Mendelian randomization can be particularly useful in disentangling complex genetic pathways (Thomas, 2005) using proteomic or metabolomic technologies to directly measure some of the intermediate metabolites or individual pharmacokinetic rate parameters. Here, the potential for confounding or established disease to distort the intermediate phenotypes is considerable, but potentially surmountable by focusing instead on the relationship between genes and intermediates, which should be immune to these problems. In a latent variables model (Conti et al., 2003), one could imagine using control samples (preferably from a cohort if time-dependent exposures are involved) to build a model for the relationship between flawed measurements of the latent phenotypes and their genetic–environmental predictors and then apply these predicted latent variables and disease in case–control comparisons.

Setting aside the particular problems inherent in a search for modifier genes, Prentice et al. provide a thoughtful discussion of some of the challenges of study design and analysis in GWASs for main effects. In particular, I would like to commend them for their discussion of the study design challenges in the use of DNA pooling and the considerable advantages of using some form of the multistage sampling design proposed by Satagopan et al. (2002), Satagopan and Elston (2003), and Satagopan, Venkatraman, and Begg (2004). The use of DNA pooling is still in its infancy, with considerable controversy about experimental designs to allow for the various sources of error in pool construction and measurement, reviewed by Prentice et al. None of the currently ongoing or proposed studies summarized by Thomas et al. (2005) have elected to use this approach, fearing loss of power to detect modest differences in allele frequencies in the face of potential measurement error, even if later stages using individual genotyping are effective in eliminating false positive signals. The WHI will be one of the first to employ this technology in the first stage of their GWAS, so its performance will be closely watched by other investigators, as the potential savings in cost could be well worth a modest loss of power if that could be overcome by using sufficiently large or sufficiently many pools. It is worth noting, however, that even individual genotyping, while extremely accurate, is not immune to measurement error. It is only recently that the statistical genetics community has turned its attention to the problem of dealing with genotyping error (Rice and Holmans, 2003; Kang, Gordon, and Finch, 2004), using techniques that have been widely used in environmental epidemiology for years (Prentice, 1982; Thomas, Stram, and Dwyer, 1993). A disturbing account by David Clayton (see online supplement to Thomas et al., 2005 for details) showed that both genotype call rates and concordance across platforms were differential by case–control status in a recent GWAS for type II diabetes, presumably because of shifts in the distribution of readings due to sample handling, implying that the software for genotype calling might need to be calibrated separately for cases and controls. For further discussion of design and analysis issues in GWAS, the interested reader is referred to several other recent reviews (Hirschhorn and Daly, 2005; Palmer and Cardon, 2005; Thomas et al., 2005; Wang et al., 2005).

As an analytical challenge, if the $p \gg n$ problem were not bad enough for a genomewide search for genetic main effects, the mind boggles in considering a search for all possible gene–gene interactions (Marchini, Donnelly, and Cardon, 2005), associations with all possible extended haplotypes (Lin, Chakravarti, and Cutler, 2004), or germline variants affecting the expression of all possible genes (Schadt et al., 2003) or proteomic patterns! Motivated in large part by such high-volume genomic tools, there has been a resurgence of interest recently in various exploratory data analysis techniques, such as classification and regression trees (CART) and multivariate adaptive regression spline (MARS) (Cook, Zee, and Ridker, 2004), neural networks (Sebastiani, Yu, and Ramoni, 2003), multifactor dimensionality reduction (Ritchie et al., 2001), random forests (Bureau et al., 2003), Bayesian

network analysis (Friedman et al., 2000) and model selection (Sillanpaa and Corander, 2002), self-organizing maps (Tamayo et al., 1999), support vector machines (Byvatov and Schneider, 2003), sequential filtering and boosting (Yasui et al., 2003; Feng, Prentice, and Srivastava, 2004), and others (Hoh and Ott, 2003) for exploring large arrays of main effects and interactions. While classical hypothesis-driven methods still have an important role to play, I expect that the future will see a greater interplay between these two basic approaches to statistical analysis of high-volume genomic data.

Proteomic methods have potentially many uses (Sellers and Yates, 2003), ranging from etiologic (e.g., dissecting complex pathways, reducing etiologic heterogeneity by subphenotyping) to clinical research and practice (e.g., early detection). The recent report of a proteomic marker for ovarian cancer (Petricoin et al., 2002) has been controversial, with the possibility of differential measurement error due to differences in sample handling among the reasons being suggested for the apparent case–control difference in proteomic patterns (Diamandis, 2004). It is an open question whether this striking difference between postdiagnostic case and control patterns will be useful as a pre-diagnostic screening test. This question will require prospective evaluation. Given the rarity of ovarian cancer, this will presumably be feasible only in high-risk cohorts, such as relatives of *BRCA1* mutation carriers.

In summary, the rapidly exploding availability of high-volume genomics technologies, including SNP genotyping, gene expression arrays, proteomics, metabolomics, and other "Omics" technologies, poses daunting challenges and opportunities that should be a high priority for statistical research.

References

Bureau, A., Dupuis, J., Hayward, B., Falls, K., and Van Eerdewegh, P. (2003). Mapping complex traits using random forests. *BMC Genetics* **4,** S64.

Byvatov, E. and Schneider, G. (2003). Support vector machine applications in bioinformatics. *Applied Bioinformatics* **2,** 67–77.

Conti, D. V., Cortessis, V., Molitor, J., and Thomas, D. C. (2003). Bayesian modeling of complex metabolic pathways. *Human Heredity* **56,** 83–93.

Cook, N. R., Zee, R. Y., and Ridker, P. M. (2004). Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Statistics in Medicine* **23,** 1439–1453.

Davey Smith, G. and Ebrahim, S. (2003). 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32,** 1–22.

Diamandis, E. P. (2004). Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations. *Molecular and Cellular Proteomics* **3,** 367–378.

Feng, Z., Prentice, R., and Srivastava, S. (2004). Research issues and strategies for genomic and proteomic biomarker discovery and validation: A statistical perspective. *Pharmacogenomics* **5,** 709–719.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7,** 601–620.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common disease and complex traits. *Nature Reviews Genetics* **6,** 95–108.

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics* **4,** 701–709.

Kahn, J. (2005). Misreading race and genomics after BiDil. *Nature Genetics* **37,** 655–656.

Kang, S. J., Gordon, D., and Finch, S. J. (2004). What SNP genotyping errors are most costly for genetic association studies? *Genetic Epidemiology* **26,** 132–141.

Klein, R. J., Zeiss, C., Chew, E. Y., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389.

Kraft, P. and Thomas, D. C. (2004). Case-sibling gene-association studies for diseases with variable age at onset. *Statistics in Medicine* **23,** 3697–3712.

Lin, S., Chakravarti, A., and Cutler, D. J. (2004). Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genetics* **36,** 1181–1188.

Little, J. and Khoury, M. J. (2003). Mendelian randomization: A new spin or real progress? *Lancet* **362,** 390–391.

Maraganore, D. M., de Andrade, M., Lesnick, T. G., Strain, K. J., Farrer, M. J., Rocca, W. A., Krishna Pant, P. V., Frazer, K. A., Cox, D. R., and Ballinger, D. G. (2005). High-resolution whole-genome association study of Parkinson disease. *American Journal of Human Genetics* **77,** 685–693.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37,** 413–417.

Need, A. C., Motulsky, A. G., and Goldstein, D. B. (2005). Priorities and standards in pharmacogenetic research. *Nature Genetics* **37,** 671–681.

Ozaki, K., Ohnishi, Y., Iida, A., et al. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* **32,** 650–654.

Palmer, L. J. and Cardon, L. R. (2005). Shaking the tree: Mapping complex disease genes using linkage disequilibrium. *Lancet* **366,** 1223–1234.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., et al. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in Cox's regression model. *Biometrika* **69,** 331–342.

Rice, K. M. and Holmans, P. (2003). Allowing for genotyping error in analysis of unmatched case-control studies. *Annals of Human Genetics* **67,** 165–174.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273,** 1616–1617.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics* **69,** 138–147.

Satagopan, J. M. and Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genetic Epidemiology* **25,** 149–157.

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E., and Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics* **58,** 163–170.

Satagopan, J. M., Venkatraman, E. S., and Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60,** 589–597.

Schadt, E. E., Monks, S. A., Drake, T. A., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297–302.

Sebastiani, P., Yu, Y. H., and Ramoni, M. F. (2003). Bayesian machine learning and its potential applications to the genomic study of oral oncology. *Advances in Dental Research* **17,** 104–108.

Sellers, T. A. and Yates, J. R. (2003). Review of proteomics with applications to genetic epidemiology. *Genetic Epidemiology* **24,** 83–98.

Sillanpaa, M. J. and Corander, J. (2002). Model choice in gene mapping: What and why. *Trends in Genetics* **18,** 301–307.

Smith, P. G. and Day, N. E. (1984). The design of case-control studies: The influence of confounding and interaction effects. *International Journal of Epidemiology* **13,** 356–365.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52,** 506–516.

Tamayo, P., Slonim, D., Mesirov, J., et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 2907–2912.

Thomas, D. C. (2005). The need for a comprehensive approach to complex pathways in molecular epidemiology (editorial). *Cancer Epidemiology, Biomarkers and Prevention* **14,** 557–559.

Thomas, D. C. and Conti, D. V. (2004). Commentary: The concept of 'Mendelian Randomization.' *International Journal of Epidemiology* **33,** 21–25.

Thomas, D. C. and Witte, J. S. (2002). Point: Population stratification: A problem for case-control studies of candidate gene associations? *Cancer Epidemiology, Biomarkers and Prevention* **11,** 505–512.

Thomas, D. C., Stram, D., and Dwyer, J. (1993). Exposure measurement error: Influence on exposure-disease relationships and methods of correction. *Annual Reviews of Public Health* **14,** 69–93.

Thomas, D. C., Haile, R. W., and Duggan, D. (2005). Design and analysis of genomewide association scans: A workshop report and review. *American Journal of Human Genetics* **77,** 337–345.

Wacholder, S., Rothman, N., and Caporaso, N. (2002). Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiologic studies of common polymorphisms and cancer. *Cancer Epidemiology, Biomarkers and Prevention* **11,** 513–520.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005). Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics* **6,** 109–118.

Yasui, Y., Pepe, M., Thompson, M. L., et al. (2003). A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* **4,** 449–463.

Anastasios A. Tsiatis and Marie Davidian

Department of Statistics
Box 8203, North Carolina State University
Raleigh, North Carolina 27695-8203, U.S.A.
email: *davidian@stat.ncsu.edu*

1. Introduction

The Women's Health Initiative (WHI) is a multifaceted public health undertaking of enormous scale involving intertwined interventional and observational components, a plethora of biological substudies, and the collection of an unparalleled data resource that will undoubtedly shape research on women's health for years to come. The entire WHI project team deserves considerable recognition for their innovative efforts in designing and implementing such an ambitious and important study.

We congratulate Drs Prentice, Pettinger, and Anderson for a thought-provoking, well-written, and comprehensive discussion of the myriad statistical research challenges posed by a study of this magnitude and importance. In addition to identifying and elucidating these challenges, the authors have explicitly highlighted the vital role of statistical science in multidisciplinary public health research, rightly emphasizing that "the statistical role (is) on a par with that of other key disciplines." We expect that this stimulating paper will inspire established and new statistical researchers alike to pursue methodological breakthroughs that will advance the scientific agenda in women's health research, disease prevention research, and clinical research more generally.

We cannot hope, nor do we feel qualified, to comment on the vast and diverse set of challenges set forth in this paper. Accordingly, we limit our remarks to the following two topics.

2. Clinical Trial Monitoring and Reporting Methods

The issue of how a Data Safety and Monitoring Board (DSMB) can assess and weigh early observed risks against

potential future benefits of a treatment is a critical one for any large-scale trial, be it in the context of disease prevention or of treatment of chronic disease. The authors' account of the preparations and procedures put in place in the WHI clinical trial (CT) anticipating such difficulty and of the high-profile early stopping of the combined hormonal trial is fascinating and highlights with clarity the complexity of the issue and the challenges facing a DSMB in this situation. As detailed by the authors, the WHI team rightly devoted considerable thought and effort prior to the CT to develop a cohesive monitoring strategy based on a combination of weighted and unweighted log-rank test statistics for not only the primary endpoints but also secondary and adverse outcomes and a "global index" determined by DSMB members' reactions to various hypothetical trial scenarios.

This account, along with our experience serving on a number of DSMBs for chronic disease trials where the early risk/future benefit conundrum has arisen, has inspired us to formulate with greater specificity an idea for trial design and analysis that we have contemplated informally for some time. This idea is meant to apply generically to trials where there is concern that early differences in the primary endpoint could emerge.

As discussed by the authors, when a time to event is the primary endpoint, the log-rank test statistic is the most common basis for monitoring chronic disease clinical trials. In the design of such trials, it is furthermore routine to assume a proportional hazards relationship between treatments. This proves convenient mathematically, as under this assumption the log-rank test statistic behaves like a Brownian motion with drift parameter related to the log-hazard ratio of interest. This allows the use of standard calculations for sequential test statistics with Brownian motion structure to develop stopping boundaries.

The difficulty, of course, is that one does not know a priori the true relationship between the hazards. Consequently, it is possible, for example, that early treatment differences may lead to a test statistic that crosses a stopping boundary with the active treatment showing an increased number of primary events (possibly deaths). This in turn leads to discussion within the DSMB of termination of the study at a time when there is potentially not a great deal of patient follow-up. This is a complex dilemma for members of a DSMB. Faced with an increased number of events on the active treatment sufficiently large to have crossed a sequential boundary, which would dictate stopping the study, the DSMB must consider the difficult issue of whether it is ethical to continue the study nonetheless with the hope that long-term benefits of the treatment may emerge, which, of course, is unknown at the time of the decision. Conversely, an early difference favoring the active treatment resulting in a boundary crossing could be observed, again before adequate patient follow-up for assessing long-term benefit. In this case, adherence to the statistical procedure would dictate stopping the study. In some instances, DSMB members may instead regard the stopping boundaries as "guidelines" rather than strict criteria and opt to continue the study in order to enhance follow-up with an eye toward assessing long-term benefit. However, this may lead to difficulties in assessing the statistical operating characteristics of the sequential test.

We believe that it may be productive to contemplate this issue at the design stage by decoupling explicitly short- and long-term effects. In particular, it may be appropriate, based on the current state of scientific knowledge, perhaps through observational evidence, and of scientific interest to focus specifically on the long-term effect of treatment. If potential long-term benefits of treatment ultimately are of interest, which, indeed, is often the case for chronic disease, then, from a statistical point of view, a parameter that characterizes long-term effect may be a better reflection of the scientific objective than the log-hazard ratio, which represents overall effect throughout time. This reasoning suggests choosing the primary endpoint to target such a parameter. For example, if ultimately long-term survival is of scientific importance, we may base the primary endpoint on the difference in survival distributions at a key point in time t_*, such as $t_* = 4$ or 5 years, and focus on the parameter

$$\delta = S_1(t_*) - S_0(t_*),$$

where $S_k(x) = P(T \geq x)$ under treatment k, $k = 0$ (placebo) or 1 (active treatment), and T is the time to event. Alternatively, we might consider area under the survival curve during some key time interval; i.e., the parameter

$$\delta = \int_{t_1}^{t_2} S_1(x)\,dx - \int_{t_1}^{t_2} S_0(x)\,dx$$

for $(t_1, t_2) = (2,5)$ years, say. Some practitioners may be reluctant to adopt such a strategy under the perception that monitoring procedures based on estimators for such parameters may be prohibitively complicated to implement. However, this is not the case; as long as an efficient estimator for the parameter is available, the information-based monitoring theory of Scharfstein, Tsiatis, and Robins (1997) leads straightforwardly to feasible techniques for sequential testing.

With such primary endpoints, because the parameter of interest cannot be estimated until sufficient follow-up is available, sequential boundaries could not be crossed early and hence one could not stop the study until a sufficient number of patients were observed for the appropriate length of time. Focusing on such primary endpoints does not eliminate the possibility that a large number of events may be seen early in the trial and the attendant ethical considerations. Under this view, monitoring for early differences in numbers of events would still take place; however, this could be considered in the same spirit as monitoring for a safety endpoint and would be separate from monitoring the primary endpoint.

Of course, as is current practice, subjective judgment could be included in considerations for decision making under early treatment differences if desired. As advocated by the authors, data from outside sources, if they exist, could be a useful supplement in determining whether such early results are sufficiently worrisome as to warrant termination of the trial.

3. Intervention Adherence and Causal Inference Methods

In the face of noncompliance to assigned treatment, the authors note that so-called adherence-adjusted analysis has been advocated, where one attempts to estimate a parameter that represents treatment effect if in fact subjects were to practice

full compliance with their assigned treatments. We agree with the authors that focusing on such a parameter can be misplaced; the authors provide compelling examples of situations in which the reasons for non-adherence are treatment related and in fact involve adverse outcomes that may preclude further treatment.

Our view is that a more feasible and realistic objective is to adopt the perspective of Murphy, van der Laan, and Robins (2001) and focus instead on identifying the effects of relevant so-called dynamic treatment regimes. For example, one might view a need to stop treatment because of adverse outcomes to be part of the overall strategy or policy of administering the treatment rather than as representing lack of compliance to the treatment. Here, then, the objective of inference would be to compare different such strategies, possibly with the ultimate goal of identifying the optimal such strategy (Murphy, 2003). A simple example is given by Johnson and Tsiatis (2004) and Rebeiz et al. (2004), who adopted this perspective in estimating the effect of treatment duration on outcome when treatment must be terminated due to an adverse event.

More generally, it is indisputable that the intention-to-treat question is of central interest and importance in randomized trials. However, the actual treatment of patients in practice is often significantly more complex than suggested by baseline treatment assignment, with numerous treatment decisions made over time. Questions involving modification of assigned treatment, including those addressing adherence issues, although arising in the context of a randomized study, are observational in nature. Formal causal inference methodology provides a systematic framework for addressing such questions. As noted by the authors, use of this framework entails critical and unverifiable assumptions, although any attempt to address these questions of necessity would involve these or related assumptions. Nonetheless, we believe that viewing such problems through the lens of causal inference methods can help to clarify the questions and formalize exactly what is required in order to address them. Accordingly, we agree with the authors that issues of adherence modeling and interpretation deserve continued development, and we believe that progress can be made by casting them as we have described. As suggested by the authors, it is essential that this be carried out through specific applications such as the WHI so that the extent and nature of assumptions required can be made transparent and hence debated in a genuine scientific context.

4. Conclusion

We again applaud the authors for an excellent paper that provides the statistical community with a rich source of methodological challenges. We are grateful for the opportunity to comment on this important piece of work.

Acknowledgements

This work was supported by NIH grants R01-CA051962, R01-CA085848, R37-AI031789, and R21-DA019800.

References

Johnson, B. A. and Tsiatis, A. A. (2004). Estimating mean response as a function of treatment duration in an observational study, where duration may be informatively censored. *Biometrics* **60,** 315–323.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B* **65,** 331–355.

Murphy, S. A., van der Laan, M. J., and Robins, J. M. (2001). Marginal models for dynamic regimes. *Journal of the American Statistical Association* **96,** 1410–1423.

Rebeiz, A. G., Derry, J. P., Tsiatis, A. A., et al. (2004). Optimal duration of eptifibatide in percutaneous coronary intervention. *American Journal of Cardiology* **94,** 926–929.

Scharfstein, D. O., Tsiatis, A. A., and Robins, J. M. (1997). Semiparametric efficiency and its implication on the design and analysis of group sequential studies. *Journal of the American Statistical Association* **92,** 1342–1350.

Rejoinder

Ross L. Prentice, Mary Pettinger and Garnet L. Anderson

We would like to thank each of the persons who provided comments on our article. Our article was intended primarily to draw attention to topics having important statistical content where further methodology development is needed. Hence, we particularly appreciate the new modeling approaches that some commentators suggested and we will provide some reaction to these. Our article also provided an update on the Women's Health Initiative and a description of the statistical methods we have employed to date in certain areas. We appreciate the critique of these methods. As we prepare to publish the principal results from the dietary modification and calcium and vitamin D components of the WHI clinical trial, and plan novel uses of the WHI specimen repository and database, this critique is like a thoughtful consultancy session, by a very high-priced group of statistical and epidemiological consultants!!

1. General Reviewer Comments

Several writers commented on the public health importance of the questions being addressed in the WHI, and on the appropriateness of the basic study design, comprising a multifaceted clinical trial and a companion cohort study. We appreciate these comments, while acknowledging that many people contributed to these and other design choices, including colleagues at NIH and at WHI clinical centers. It is also worth commenting that this type of enterprise is not at all

easy to initiate. Investigators around the country, both outside and within NIH, worked for years in attempts to launch full-scale clinical trials of a low-fat eating pattern and of postmenopausal hormone therapy. Partly because of the cost of such trials, there is typically a resistance to such proposals from the research community. It is perhaps unlikely that either of these trials would have been launched had not Dr Bernedine Healy sought and received the necessary funding as a special congressional appropriation shortly after assuming the NIH director position. We would also like to comment that study of the research designs needed to obtain reliable public health information, the infrastructure needs for a vibrant preventive intervention development program, and the need for suitable forums to promote the needed research and to advise funding organizations on topics where full-scale trials may be warranted, are extremely important issues for public health research progress, but are mostly beyond the scope of our article. See Prentice et al. (2004) for a discussion of these topics in the diet and physical activity epidemiology research areas.

Several commentators also resonated with our remarks about biostatistics being one of the fundamental disciplines in the public health research arena. As illustrated by the topics we chose to emphasize, methodologic issues are often crucial to progress in public health–oriented research. Our colleagues in areas such as clinical trials, epidemiology, nutrition, and genetics do excellent work on both substantive and methodologic topics, but statistically oriented investigators are often needed to help identify the research gaps and to develop sound quantitative approaches to filling these gaps. The commentator group includes excellent role models who have greatly impacted such research areas as nutritional epidemiology, genetic epidemiology, and clinical trial methodology, to name just a few.

2. Nutritional Epidemiology Methods

Turning to comments on dietary measurement error modeling and related topics, Dr Raymond Carroll, who has been an important contributor to this area, emphasizes the distinction between recovery biomarkers, which typically arise from urinary measures that reflect the body's expenditure of a nutrient, and concentration biomarkers, which assess circulating levels of a nutrient in blood or other body compartments. The former measures may plausibly adhere to a classical measurement model, but are available for only a few nutrients, while the latter are available for many more nutrients, but will typically include person-specific biases that make them generally unsuitable for the purpose of calibrating dietary self-report estimates. Dr Carroll asks our opinion about two potential ways of using concentration biomarkers. The first involves focusing on disease association with the biomarker, rather than the nutrient consumed. This approach simplifies the analysis, and yields association parameters of interest, but the resulting information would not seem directly useful for the development of evidence-based dietary pattern recommendations. His second approach to using concentration biomarkers involves human feeding studies. There are rather few research groups configured for human feeding trials or exercise intervention trials. Nevertheless, we think this approach could be feasible and useful. For example, a study in which individual study subjects are exposed to two or more known dietary intakes for a nutrient and corresponding (lagged) blood concentrations are recorded, could be used to estimate the parameters in Dr Carroll's model (3). One might then transfer, for example, the estimated measurement error correlation, derived from concentration marker measures on the same individual, from the feeding study to a cohort or clinical trial setting. Doing so could allow a separation of the (unknown) nutrient consumption from the person-specific bias in Dr Carroll's model (3) for self-report data calibration and disease association testing.

Dr Nick Day also brings a wealth of experience to the nutritional epidemiology and biomarker development areas. He comments that diet and physical activity patterns appear to be key determinants of a range of health outcomes, and notes the extraordinary difficulty of related epidemiologic research. He asks very pertinent and timely questions concerning the ability to interpret the WHI dietary modification trial results, which are currently being prepared for publication, given the major uncertainties that attend the assessment of diet, and the dietary change induced by intervention.

Fundamentally, the DM trial assesses whether a dietary intervention program having certain goals (20% of energy from fat, 5 or more servings of fruits and vegetables per day, ...), applied in a certain manner (nutritionist-led small group sessions having nutritional and behavioral content), can reduce the incidence of breast cancer, colorectal cancer, and the other designated outcomes over an average 8.1-year study period. As Dr Day suggests, a positive answer to this question will greatly advance public health aspects of nutrition, in spite of a limited ability to attribute the disease risk reduction to the specific dietary changes made (or to other changes that intervention group women may have chosen to a different extent than did control group women). If, however, a null or equivocal result arises, then trial interpretation may well depend on the extent of dietary differences between the intervention and control groups, and the ability to assess nutrient intakes, even at the group level, may require suitable biomarkers.

We agree with Dr Day that current technology does not allow a compelling response to these issues. The largest problem with the food frequency questionnaire, which is our basic tool for measuring dietary change in WHI, concerns total energy assessment. Our nearly completed Nutrient Biomarker Study will permit a calibrated estimate of total energy at baseline and at various subsequent time points, with a calibration procedure that depends on intervention group assignment. These calibrated estimates can be combined with FFQ percent energy from fat estimates to obtain total fat consumption estimates, and changes in total fat consumption that are expected to be an improvement over FFQ fat consumption estimates. We will examine the extent that these estimates of change mediate any intervention effect on disease outcome. The biomarker data will also allow a calibrated assessment of nonprotein calorie consumption. We have an interest in using the respiratory quotient from resting state indirect calorimetry in an attempt to separate (even if noisily) fat and carbohydrate consumption, but necessary funding has yet to be secured for this work. Another aspect of DM trial interpretation concerns study of the relationship between any observed

intervention effect and baseline dietary habits. Baseline FFQ nutrient consumption estimates are distorted by our use of the FFQ as a screening instrument (≥32% energy from fat for eligibility). To reduce our reliance on FFQ, we also plan to present analyses as a function of nutrient consumption based on four-day food records, using a case-only analysis.

In response to Dr Sander Greenland's nutritional epidemiology comments, we note that both the self-report assessments and the biomarker assessments typically pertain to a short time period of a few days to a few months. In the WHI dietary modification trial, for example, we sought food frequency data at baseline and 1 year from randomization on all women, and subsequently, approximately every 3 years in a rotating sample basis. Our biomarker calibration equations derive from the FFQ and biomarker correspondence at about 8 years from randomization. These equations will be applied to the various FFQs collected for a woman, to give a biomarker-corrected dietary history over the average 8.1-year trial follow-up period.

3. Genomic and Proteomic Methods

We appreciate Dr Duncan Thomas's comments on the challenges in detecting genetic modifiers of the effect of a treatment or intervention, reflecting his many years of contributions to the genetic epidemiology literature. Dr Thomas provides very up-to-date citations of initial reports from genome-wide association studies, as well as citations to recent articles discussing related methodologic issues. Since the time of writing our article we have proceeded with the early implementation of a genome-wide association study in collaboration with Dr David Cox (no, not that David Cox!) of Perlegen Sciences in Mountain View, California. For each of coronary heart disease, stroke, and breast cancer, this study will involve 250,000 SNPs and 1000 cases and 1000 pair-matched controls drawn from the WHI observational study in the first stage of a three-stage design. This first stage will involve eight pairs of DNA pools, each comprising equal volumes of DNA from 125 cases or 125 pair-matched controls. DNA from racial/ethnic minority cases will be placed in a single pool, and matching factors will include ethnicity as well as age, WHI enrollment date (to control for follow-up duration), and selected other factors. SNPs meeting a 1% significance level criterion for either a test based on an allele frequency difference statistic or an odds ratio statistic will move on to the second stage, which will involve individual SNP determinations for about 613–800 cases and controls (depending on the disease) also drawn from the observational study. SNPs meeting a 2% significance level criterion at this stage will be examined individually in estrogen plus progestin clinical trial cases and controls (258–349 cases and controls) with testing at the 5% significance level. We have conducted simulation studies (Ross Prentice and Lihong Qi, submitted for publication, 2005) to indicate that such a design can be expected to have adequate power for detecting an odds ratio of 1.5 or greater for the minor SNP allele, provided the allele frequency is not too small (e.g., ≥0.2) and provided an additive or dominant genetic model prevails, with lesser power under a recessive model. Because this design implies an overall significance level of $(0.01)(0.02)(0.05) = 0.00001$, there will be only 2.5 expected false positives under the global null hypothesis, facilitating decision making as to whether disease-related SNPs have been identified. We also found, in these simulation studies, that a test statistic that combines data across the preceding stages with testing at 0.01, 0.0002, and 0.00001 levels, respectively, has improved power properties compared to a separate testing procedure at each design stage.

Both Dr Thomas and Dr David DeMets point to the need for statistical methodology development for the design and analysis of genome-wide association studies. In fact, the National Heart, Lung, and Blood Institute and several other NIH Institutes recently issued a request for application (RFA-HL-05-011) precisely to encourage this type of methodology development. The RFA calls for initiatives on a number of topics including tagging SNP selection, assessment of the utility of pooled DNA, study design and analysis choices, haplotype block formation methods, and methods for gene–gene or gene–environment interaction, and refers to the absence of suitable methods as the key bottleneck for the progress in this promising research area, given the impressive technical advances in high-throughput SNP genotyping of the past few years.

Dr Greenland finds it "very odd" that we "neglected" empirical Bayes and other shrinkage procedures in our discussion of this high-dimensional genomic and proteomic studies area. We find his criticism very odd in that our presentation involved only testing and identification procedures, and did not discuss any form of estimation procedure. Of course, there are interesting estimation methods issues in the context of the type of multistage studies mentioned above. For example, one may be interested in estimating the disease odds ratio associated with an SNP or a haplotype block that acknowledges both the high dimensionality and the selection procedure employed (e.g., Benjamini and Yekutieli, 2005). These methods will require some form of shrinkage for parameter estimation. The interesting article (Efron, 2004) that Dr Greenland cites is concerned with the choice of the null hypothesis for the high-dimensional testing problem. Our own plans for methodology development, included in responses to the RFA just mentioned, include an empirical comparison of the efficiency of empirical versus theoretical null hypothesis testing procedures in the WHI–Perlegen data, as well as an exploration of shrinkage options for parameter estimation.

4. Clinical Trial Monitoring Procedures

Drs DeMets and Day offer rather different views of the adequacy of prevailing clinical trial monitoring methods in complex settings such as the WHI, where study treatments or interventions plausibly affect multiple important clinical outcomes, each with its own time course and severity. Dr DeMets provides interesting insights into the deliberations of the WHI DSMB, some of which we, as WHI investigators, were not a party to. He describes aspects of the clinical trial monitoring plan developed by WHI-related statisticians in collaboration with the DSMB and concludes that, for the hormone therapy trials, the global index we defined as a supplementary statistic to that for the primary outcomes was "not as useful as originally intended" since the outcomes included in the global index were going in different directions. Dr DeMets comments further that "no additional statistical methodology" would have facilitated DSMB recommendations.

Dr Day, in contrast, finds it "troubling" that early results had such an influence in triggering the early stopping of the

hormone therapy trials. He offers the perspective that these trials, and others such as the breast cancer prevention trial of tamoxifen, should have been allowed to continue "sufficiently to generate data of unambiguous value for clinical or public health decisions," and he concurs with our call for the formulation of stopping rules that "provide a more helpful balance between short- and longer-term effects."

While it is clear that monitoring committees need to be in a position to exercise judgment beyond those implied by formal statistical monitoring procedures in such complex situations, we share with Dr Day the viewpoint that premature stopping is a serious concern in prevention trial monitoring. The development of more flexible and comprehensive statistical monitoring procedures seems to be a major tool toward realizing the scientific potential of prevention trials, while taking suitable account of ethical issues. Our own experience, both as members of monitoring committees and as recipients of their recommendations, suggests that structure is needed, in spite of committee member expertise and experience, since the exposure of members to trial data is typically brief and episodic, and since the reaction of members to trial data tends to be highly variable and dependent on personal research interests and perceptions of committee responsibilities.

Even when formal monitoring guidelines have been adopted, it is difficult for monitoring committees and sponsors to allow a trial to continue in the face of data indicating early harm. For example, in the WHI estrogen-only trial, the global index was almost exactly balanced between benefits and risks when the early stopping occurred. This type of situation, rather than a situation in which all the pertinent outcomes were going in the same direction, was the motivation for the inclusion of the global index in the monitoring plan, that is, to help prevent premature stopping if there is major uncertainty concerning overall benefits versus risks at a particular point in time in trial conduct.

The WHI trial monitoring plan may not have included sufficient provision for a changing course of benefits versus risks as the time from the initiation of treatment increases. As Dr Anastasios Tsiatis and Dr Marie Davidian point out, one usually does not know in advance of trial conduct the form of the treatment hazard ratio (HR) over time, and it may turn out that the proportional hazards assumption that underlies most statistical monitoring procedures is far from correct, as for coronary heart disease, venous thromboembolism, and breast cancer in the WHI estrogen-plus-progestin trial.

Hence, we concur with Drs Tsiatis and Davidian that monitoring procedures that attempt to disassociate short-term from long-term effects could be quite useful. One needs only to consider a trial of a surgical intervention having some short-term mortality along with putative longer-term benefit to realize that some form of disassociation may be essential to avoid certain premature stopping in a sufficiently large trial.

In response to the interesting specific proposals of Drs Tsiatis and Davidian, it seems to us that it may not be desirable to go so far as to relegate the early data on a primary outcome to a separate adverse events monitoring status, since an early effect does need to be an element of a primary endpoint benefit versus risk summary, over any specified treatment/follow-up period. Our own attempts to address this issue involved weighted log-rank statistics, with weights increasing with time from randomization. With this type of statistic the influence of the early data declines as longer-term data accumulate. In the case of the estrogen-plus-progestin trial, the breast cancer weighted log-rank statistic crossed a monitoring boundary, and the global index statistic also met adverse effects criteria sufficient to support an early trial-stopping consideration. This configuration, along with data on a number of other clinical outcomes that were monitored informally, seems to us to provide an answer to the public health question about the balance of risks and benefits in the study population over the 5.6-year average follow-up period. While it would be helpful to know longer-term effects (as well as effects in important subsets), the essential question addressed by this trial as to whether hormone therapy could be advocated in terms of benefits for the primary endpoint coronary heart disease and in terms of overall health benefits versus risks in populations like WHI had been substantially answered, and we have no argument with our DSMB concerning their early stopping recommendation. As Dr Day implied, these trial results were difficult for certain practicing and research communities to accept, though we see this more as a result of overinterpretation of the preceding observational study data than as being due to the absence of longer-term clinical trial data. The estrogen-alone trial, on the other hand, even though having a longer average follow-up time (7.1 years), seems to us to leave greater uncertainty concerning clinical and public health recommendations, as the stroke and venous thromboembolism elevations are offset by fracture reductions, a possible breast cancer reduction, and even a suggestion of a favorable trend in coronary heart disease toward the end of the trial. These are complex issues to assimilate and there seems to us to be a need for additional development and application of monitoring procedures that can adapt to HR changes over time for several outcomes, and for corresponding estimation procedures that acknowledge such complex monitoring systems.

5. Postmenopausal Hormone Therapy and Cardiovascular Disease

We appreciate the vigorous discussion of our combined analysis of WHI clinical trial and observational data on estrogen plus progestin and cardiovascular disease. As Dr Day describes, various commentators in other forums have pulled out various theories, which Dr Day describes as "empty rhetoric," to explain this discrepancy. We do think that the availability in WHI of clinical trial and observational study data drawn from the same population, over the same time period, using essentially the same data collection instruments, provides a strong setting to examine this discrepancy. Dr David Freedman and Dr Diana Petitti note that "without experimental data, it will be unclear which adjustments to make, or how far to go."

In analyses too detailed to present in Prentice et al. (2005) or in the present article, we examined various potential additional sources of confounding. These included interactions among established disease risk factors, treatment–covariate interactions (Dr Greenland's question), changing risk factors across follow-up time, and classical measurement error "corrected" risk factors in the diet and physical activity areas. While it would be an overstatement to say that these analyses were exhaustive, we were impressed at the insensitivity

of hormone therapy HRs in the OS, and comparative HRs in the CT to OS to these refinements.

It is quite possible, however, as Drs Freedman and Petitti argue, that there is some residual confounding in the OS HR estimate for both coronary heart disease and venous thromboembolism, as is almost certainly the case for stroke (which we chose not to include for reasons of space; but see Prentice et al., 2005). We speculate that residual confounding in this setting is likely due to factors that are simply not being entertained, in spite of the unusually rich WHI database, or possibly due to the inability to adequately correct for important, but poorly measured dietary, physical activity, and other behavioral factors.

Drs Freedman and Petitti argue that, because of limited power to compare clinical trial and observational study HRs, their "null hypothesis" that the WHI observational study underestimates hormone therapy risks by a factor of 1.5 to 3 is as compatible with the WHI data as is a hypothesis of equality between WHI clinical and observational study risks. Our recent article (Prentice et al., 2005), which appeared with an invited commentary by Drs Petitti and Freedman, provided hazard estimates (95% confidence intervals) for the ratio of estrogen-plus-progestin HRs in the OS to those in the CT, following adjustment for the confounding factors listed and with time-from-initiation accommodation. These were 0.93 (0.64, 1.36) for coronary heart disease and 0.84 (0.54, 1.29) for venous thromboembolism—consistent with equality even though noisy, but hardly consistent with a 3-fold underestimation in the observational study! Even for stroke the corresponding numbers were 0.76 (0.49, 1.18). We would hasten to add, however, that our purpose was not to argue that observational data, carefully analyzed, can somehow obviate the need for clinical trial data. As we have noted in other places (e.g., Prentice et al., 2004), we think there is a need for organizational strengthening in the public health research community so that treatments or interventions having the most compelling rationale and public health potential are identified and receive the support of this community for evaluation in randomized controlled trials. Such structure is particularly needed in view of typical high research costs, and frequent relevance of a broad range of health outcomes that may, for example, cut across the foci of several NIH institutes. We disagree somewhat with Drs Freedman and Petitti when they "suggest that observation necessarily precedes experiment." In our view, a vigorous prevention research program may well include interventions that are self-selected by few, if any, persons in populations of interest, in which case hypothesis development and initial testing may be primarily based on small-scale trials (e.g., dietary or physical activity trials) having intermediate outcome endpoints, and on related basic science research. Even when sizeable numbers of persons adopt a dietary or other lifestyle factor having preventive potential, it may be that observational studies are not sufficiently reliable to provide much guidance to the research enterprise. As Dr Day points out, "negative results can be at least as suspect as positive ones" in nutritional epidemiology settings.

We appreciate the nice contribution by Drs Miguel Hernán, James Robins, and Luis García Rodríguez (HRGR), which provides analysis of data from the General Practice Research Database (GPRD) on estrogen plus progestin in relation to coronary heart disease. As they mention, they originally intended to include an analysis of Nurses, Health Study cohort data as well, which stimulated us to include some corresponding comments and sensitivity analyses near the end of our article. Specifically, we noted that the Nurses' Health Study presentations relied on biennial snapshots of hormone therapy exposure, and we drew upon patterns of starting and stopping hormone therapy in community studies to illustrate that an early adverse effect as observed in the WHI trial could easily be missed with the data and analytic approaches used to date in the Nurses' Health Study. This illustrates that the reasons for bias in observational studies can be diverse and also specific to a study environment. We will leave it to the reader to assess whether we have revealed ourselves to be closet Bayesians for having used these community data to inform our little simulation experiment, as Dr Greenland appears to suggest.

As defined by HRGR, the GPRD cohort includes 99,072 women of whom 1889 experienced CHD during follow-up. Apparently, a substantial number of these women initiated estrogen-plus-progestin therapy during the defined follow-up period, motivating their "intent-to-treat" (ITT) analyses of these data, but only 64 CHD outcomes occurred among the initiators (compared to 188 among women randomized to estrogen plus progestin in the WHI trial) so precision is limited. Their ITT estimated HRs (95% CIs) as a function of 0–2, 2–5, and >5 years from initiation of estrogen plus progestin (their Table 2) are 1.20 (0.84, 1.72), 0.82 (0.55, 1.21), and 0.69 (0.38, 1.25), which do not seem particularly discrepant with corresponding WHI trial HRs of 1.68 (1.15, 2.45), 1.25 (0.87, 1.79), and 0.66 (0.36, 1.21), and not at all discrepant with our WHI observational study HRs of 1.12 (0.46, 2.74), 1.05 (0.70, 1.58), and 0.83 (0.67, 1.01). These last HR estimates were, for simplicity, based on estrogen-plus-progestin usage at enrollment into the observational study since there were rather few hormone treatment initiators following cohort enrollment and we chose to address their influence on HR estimation as an element of our adherence sensitivity analyses.

We think the HRGR formulation of a data analysis approach that attempts to emulate an ITT analysis in a randomized controlled trial has attractive features. It would be a mistake, however, to think that this formulation leads to results of similar reliability to a randomized trial, largely because of the strong no unmeasured confounder assumption described by HRGR. In fact, to render the treated and untreated groups comparable for a causative inference, it is not only necessary that there are no unmeasured confounders, but the potential confounders need to be accurately measured (or appropriately adjusted for measurement error), and confounder relationship to outcome and treatment must be properly modeled. As Drs Freedman and Petitti note, observational data alone do not provide much guidance as to when there is a sufficient control for confounding. We do not find HRGR's comparison of HRs adjusted for and without adjustment for available confounding factors to be at all compelling as a guide to whether or not substantial residual confounding remains. In fact, this criterion would seem to favor fewer confounding factors and less careful modeling since then adjusted and unadjusted HRs would be more similar.

HRGR's analysis evidently regards a woman as an initiator each time that she starts or restarts estrogen-plus-progestin usage. We cannot tell from their presentation how long a usage gap was required before regarding a repeat user as a new initiator. Also, the apparent time-dependent relationship between HR and duration of use would seem to imply a need for an exquisite modeling of disease risk (HRGR model (1)) on prior hormone therapy usage pattern. Our WHI observational study analyses focused on the current estrogen-plus-progestin episode at baseline, with a prior usage gap of 1 year or more determining a new episode. This allowed simple analyses that excluded women who used these preparations prior to their baseline episode and excluded clinical trial women who used them prior to randomization. We do not claim that our simple methods for addressing these issues make optimal use of WHI data, and we suspect that HRGR made reasonable corresponding modeling choices as well. Of course, in addition to modeling the time course of prior estrogen-plus-progestin use, there are issues of dose levels, schedules, agents, and routes of administration that observational studies analyses of estrogen-plus-progestin treatment also need to address.

Before leaving the ITT type of analysis, we will respond to a couple of other issues mentioned in relation to our hormone therapy and cardiovascular disease analyses. Dr Greenland assails us for an "exclusive focus on HRs." While HR estimates are an important data summary tool, and one that may be particularly valuable for comparing intervention effects across studies, we agree that HRs do not tell the complete story, especially concerning clinical and public health implications. We refer Dr Greenland to the primary estrogen-plus-progestin and estrogen-alone results articles (*Journal of the American Medical Association*, 2002, 2004) for absolute risk estimates from these trials and for estimates of numbers of events per 1000 person-years that may be added or subtracted by use of the preparations studied in a population like WHI. Drs Freedman and Petitti raise the policy issues that WHI data are not available to them, as taxpayers, for this reanalysis. We can report that a rather comprehensive data set from the estrogen-plus-progestin trial will be available to requestors through NHLBI around the end of 2005. Such a "limited access data set" can be analyzed under the auspices of your local IRB with the requirement that resulting manuscripts need to be submitted to NHLBI for review and comment prior to publication. The 3-year period between the initial trial publication and the availability of this data set is about the minimum needed for investigators and sponsors to publish principal trial findings. A similar data set from the estrogen-alone trial can be expected in 2007.

Turning now to our final topic of adherence-adjusted analyses, even a rigorously conducted randomized trial typically does not permit valid comparisons among persons having similar treatment adherence patterns, since additional untestable assumptions are needed before it can be claimed that persons in the same adherence strata are comparable in terms of baseline risk for targeted outcomes. Adopting, instead, the "dynamic treatment regimen" concept recommended by Drs Tsiatis and Davidian has a practical appeal.

We used a rather simple approach to adherence sensitivity analyses in the hormone therapy trials, censoring the follow-up time for a woman 6 months after she ceased to take 80% or more of study pills. We find it necessary to frequently remind our WHI colleagues that these adherence-adjusted HRs are not to be taken very seriously, since adherence in active and placebo groups likely differ in important disease risk-related manners as the trial progresses. Dr DeMets notes that this adherence bias issue constitutes an important advantage for trials in view of the basic ITT data analysis option. Drs Freedman and Petitti argue that adherence bias may be associated with a cardiovascular HR bias factor of 2, based on placebo group comparisons between adherers and nonadherers in prior clinical trials. Dr DeMets also offers evidence that covariate correction is unlikely to restore the desired comparability.

HRGR offer inverse probability-weighted estimation procedures to address the adherence bias issue. The method requires that women who continue or discontinue their treatment at each follow-up time are comparable conditional on available potential confounding factors for discontinuation. We agree with their approach in the sense that imposing this type of assumption may be the best one can do with available data, but as Drs Freedman and Petitti note, "adjustment must be incomplete, because relevant lifestyle factors are extraordinarily difficult to identify and measure." Baseline characteristics would not plausibly be sufficient for this purpose in complex settings, and the use of posttreatment initiation variables raises additional issues of overcorrection if adherence adjustment factors are on a pathway linking treatment to the targeted diseases. Nevertheless, we appreciate HRGR's comments on, and illustration of, the IPW approach and concur that it would be of interest to try these on the WHI hormone therapy data. This would also be the case for the associated augmented estimating equation approach, with its so-called double robust property. Our own simulation studies of these augmented estimating equations, and applications to other WHI data (Qi et al., 2005), attest to their desirable properties, provided critical modeling assumptions are met.

HRGR conclude their commentary by asking us whether we would have reached the same conclusions concerning estrogen-plus-progestin effects on cardiovascular disease as in our paper if only the observational data had been available. Our answer is "probably not." We would have been unable to detect a trend in HR as a function of time-from-initiation based on observational data alone. Our efforts to control confounding would likely have been more limited without having the trial data to help develop the HR model. HRGR also ask for our comments on the results of other observational studies. We have offered a potentially important source of discrepancy between the Nurses' Health Study and WHI trial; Drs Freedman and Petitti offer an additional suggestion stemming from lack of control in NHS for socioeconomic factors. Our companion publication (Prentice et al., 2005) offers brief suggestions in relation to other observational studies. HRGR also ask how sure are we that we have found "clear and convincing explanations." We take some comfort in the fact that persons with a wealth of related experience find our arguments convincing. For example, Dr Day writes that after examining results as a function of time-from-initiation (and confounding), "the apparent discrepancy simply disappears," while Drs Freedman and Petitti wrote that we provide "answers" that they find "persuasive." On the other hand, we agree with HRGR and Drs Freedman and Petitti that precision is limited and it

is quite possible that residual confounding of the magnitude listed by HRGR remains, again attesting to the importance of randomized controlled trials when the public health importance is great.

References

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters (with discussion). *Journal of the American Statistical Association* **100,** 71–93.

Prentice, R. L., Willett, W. C., Greenwald, P., et al. (2004). Nutrition and physical activity and chronic disease prevention: Research strategies and recommendations. *Journal of the National Cancer Institute* **96,** 1276–1287.

Prentice, R. L., Langer, R., Stefanick, M. L., et al. (2005). Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *American Journal of Epidemiology* **162,** 1–11.

Qi, L., Wang, C. Y., and Prentice, R. L. (2005). Weighted estimates of proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, in press.