

Observational Studies Analyzed Like Randomized Experiments

An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease

Miguel A. Hernán,^{a,b} Alvaro Alonso,^c Roger Logan,^a Francine Grodstein,^{a,d} Karin B. Michels,^{a,d,e} Walter C. Willett,^{a,d,f} JoAnn E. Manson,^{a,d,g} and James M. Robins^{a,h}

Background: The Women's Health Initiative randomized trial found greater coronary heart disease (CHD) risk in women assigned to estrogen/progestin therapy than in those assigned to placebo. Observational studies had previously suggested reduced CHD risk in hormone users.

Methods: Using data from the observational Nurses' Health Study, we emulated the design and intention-to-treat (ITT) analysis of the randomized trial. The observational study was conceptualized as a sequence of "trials," in which eligible women were classified as initiators or noninitiators of estrogen/progestin therapy.

Results: The ITT hazard ratios (HRs) (95% confidence intervals) of CHD for initiators versus noninitiators were 1.42 (0.92–2.20) for the first 2 years, and 0.96 (0.78–1.18) for the entire follow-up. The ITT HRs were 0.84 (0.61–1.14) in women within 10 years of menopause, and 1.12 (0.84–1.48) in the others (*P* value for interaction = 0.08). These ITT estimates are similar to those from the Women's Health Initiative. Because the ITT approach causes severe treatment misclassification, we also estimated adherence-adjusted effects by inverse probability weighting. The HRs were 1.61 (0.97–2.66) for the first 2 years, and 0.98 (0.66–1.49) for the entire follow-up. The HRs were 0.54 (0.19–1.51) in women within 10 years after menopause, and 1.20 (0.78–1.84) in others (*P* value for interaction = 0.01). We

also present comparisons between these estimates and previously reported Nurses' Health Study estimates.

Conclusions: Our findings suggest that the discrepancies between the Women's Health Initiative and Nurses' Health Study ITT estimates could be largely explained by differences in the distribution of time since menopause and length of follow-up.

(*Epidemiology* 2008;19: 766–779)

Causal inferences are drawn from both randomized experiments and observational studies. When estimates from both types of studies are available, it is reassuring to find that they are often similar.^{1–3} On the other hand, when randomized and observational estimates disagree, it is tempting to attribute the differences to the lack of random treatment assignment in observational studies.

This lack of randomization makes observational effect estimates vulnerable to confounding bias due to the different prognosis of individuals between treatment groups. The potential for confounding may diminish the enthusiasm for other desirable features of observational studies compared with randomized experiments—greater timeliness, less restrictive eligibility criteria, longer follow-up, and lower cost. However, even though randomization is the defining difference between randomized experiments and observational studies, further differences in both design and analysis are commonplace. As a consequence, observational-randomized discrepancies cannot be automatically attributed to randomization itself.

In this paper we assess the extent to which differences other than randomization contribute to discrepant observational versus randomized effect estimates in the well-known example of postmenopausal estrogen plus progestin therapy and the risk of coronary heart disease (CHD). Specifically, we explore discrepancies attributable to different distributions of time since menopause, length of follow-up, and analytic approach.

The published findings on this topic can be briefly summarized as follows. Large observational studies suggested a reduced risk of CHD among postmenopausal hormone users.

Submitted 11 January 2008; accepted 30 May 2008.

From the ^aDepartment of Epidemiology, Harvard School of Public Health, Boston, Massachusetts; ^bHarvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts; ^cDivision of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, Minnesota; ^dChanning Laboratory, Department of Medicine and ^eObstetrics and Gynecology Epidemiology Center, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; ^fDepartment of Nutrition, Harvard School of Public Health, Boston, Massachusetts; ^gDivision of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; and ^hDepartment of Biostatistics, Harvard School of Public Health, Boston, Massachusetts.

This work was supported by National Institutes of Health grants HL080644 and CA87969.

Editors' note: Commentaries on this article appear on pages xxx, xxx, xxx. Correspondence: Miguel Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA. E-mail: miguel_hernan@post.harvard.edu.

Copyright © 2008 by Lippincott Williams & Wilkins

ISSN: 1044-3983/08/1906-0766

DOI: 10.1097/EDE.0b013e3181875e61

Two of the largest observational studies were based on the Nurses' Health Study (NHS)^{4,5} in the United States and on the General Practice Research Database⁶ in the United Kingdom. More recently, the Women's Health Initiative (WHI) randomized trial⁷ found a greater incidence of coronary heart disease among postmenopausal women in the estrogen plus progestin arm than in the placebo arm (68% greater in the first 2 years after initiation, 24% greater after an average of 5.6 years).^{8,9}

The present paper does not address the complex clinical and public health issues related to hormone therapy, including risk-benefit considerations. Rather, we focus on methodologic issues in the analysis of observational cohort studies. Specifically, we reanalyze the NHS observational data to yield effect estimates of hormone therapy that are directly comparable with those of the randomized WHI trial except for the fact that hormone therapy was not randomly assigned in the NHS. We do this by mimicking the design of the randomized trial as closely as possible in the NHS. As explained below, our approach requires conceptualizing the observational NHS cohort as if it were a sequence of nonrandomized trials. Because the randomized trial data were analyzed under the intention-to-treat (ITT) principle, we analyze our NHS trials using an observational analog of ITT (see below).

A recent reanalysis of the General Practice Research Database using this strategy could not adjust for lifestyle factors and it yielded wide confidence intervals (CI).¹⁰ Further, the estrogen used by women in that study was not the conjugated equine estrogen used by the women in the NHS and WHI studies. Our analysis of the NHS data incorporates lifestyle factors and includes women using the same type of estrogen as in the WHI randomized trial.

METHODS

The Observational Cohort as a Nonrandomized "Trial"

The NHS cohort was established in 1976 and comprised 121,700 female registered nurses from 11 US states, aged 30 to 55 years. Participants have received biennial questionnaires to update information on use, duration (1–4, 5–9, 10–14, 15–19, 20–24 months), and type of hormone therapy during the 2-year interval. Common use of oral estrogen plus progestin therapy among NHS participants began in the period between the 1982 and the 1984 questionnaires. The questionnaires also record information on potential risk factors for and occurrence of major medical events, including CHD (nonfatal myocardial infarction or fatal coronary disease). The process for confirming CHD endpoints has been described in detail elsewhere.⁴

We mimicked the WHI trial by restricting the study population to postmenopausal women who in the 1982 questionnaire had reported no use of any hormone therapy during the prior 2-year period ("washout" period), and in the 1984 questionnaire reported either use of oral estrogen plus progestin

therapy ("initiators") or no use of any hormone therapy ("non-initiators") during the prior 2-year period. Thus, as in the WHI, the initiator group includes both first-time users of hormone therapy and reinitiators (who stopped hormone therapy in 1980 or earlier and then reinitiated use in the period 1982–1984).

Women were followed from the start of follow-up to diagnosis of CHD, death, loss to follow-up, or June 2000, whichever occurred first. Unlike in the randomized WHI and the observational General Practice Research Database, the time of therapy initiation—and thus the most appropriate time of start of follow-up for initiators—was not known with precision in the NHS, and so we needed to estimate it. For women who reported hormone therapy initiation during the 2-year period before the 1984 questionnaire and were still using it at the time they completed this questionnaire, the start of follow-up was estimated as the month of return of the baseline questionnaire minus the duration of hormone therapy use (duration is reported as an interval, eg, 20–24 months; we used the upper limit of the interval, eg, 24 months). For women who reported starting hormone therapy during the same 2-year period but had stopped using it by the time they returned the 1984 questionnaire, the start of follow-up was estimated as the first month of the 2-year period (the earliest possible month of initiation). The start of follow-up for noninitiators was estimated as the average month of start of follow-up among initiators (stratified by age and past use of hormone therapy). Alternative methods to estimate the start of follow-up had little effect on our estimates (Appendix A1).

To further mimic the WHI, we restricted the study population to women who, before the start of follow-up, had a uterus, no past diagnosis of cancer (except nonmelanoma skin cancer) or acute myocardial infarction, and no diagnosis of stroke since the return of the previous questionnaire. To enable adjustment for dietary factors, we restricted the population to women who had reported plausible energy intakes (2510–14,640 kJ/d) and had left fewer than 10 of 61 food items blank on the most recent food frequency questionnaire before the 1984 questionnaire.

The NHS cohort study can now be viewed as a nonrandomized, nonblinded "trial" that mimics the eligibility criteria, definition of start of follow-up, and treatment arms (initiators vs. noninitiators) of the WHI randomized trial, but with a different distribution of baseline risk factors (eg, lower age and shorter time since menopause in the NHS compared with the WHI). We analyzed the NHS nonrandomized "trial" by comparing the CHD risk of initiators and noninitiators regardless of whether these women subsequently stopped or initiated therapy. Thus our analytic approach is the observational equivalent of the ITT principle that guided the main analysis of the WHI trial. Specifically, we estimated the average hazard (rate) ratio (HR) of CHD in initiators versus noninitiators, and its 95% CI, by fitting a Cox proportional hazards model, with "time since beginning of follow-up" as the time variable,

that included a non time-varying indicator for hormone therapy initiation. The Cox model was stratified on age (in 5-year intervals) and history of use of hormone therapy (yes, no).

To obtain valid effect estimates in a nonrandomized trial, all baseline confounders have to be appropriately measured and adjusted for in the analysis. We proceeded as if this condition was at least approximately true in the NHS nonrandomized “trial” once we added the following covariates to the Cox model: parental history of myocardial infarction before age 60 (yes, no), education (graduate degree: yes, no), husband’s education (less than high school, high school graduate, college, graduate school), ethnicity (non-Hispanic white, other), age at menopause (<50, 50–53, >53), calendar time, high cholesterol (yes, no), high blood pressure (yes, no), diabetes (yes, no), angina (yes, no), stroke (yes, no), coronary revascularization (yes, no), osteoporosis (yes, no), body mass index (<23, 23–<25, 25–<30, ≥30), cigarette smoking (never, past, current 1–14 cigarettes per day, current 15–24 cigarettes per day, current ≥25 cigarettes per day), aspirin use (nonuse, 1–4 years, 5–10 years, >10 years), alcohol intake (0, >0–<5, 5–<10, 10–<15, ≥15 g/d), physical activity (6 categories), diet score (quintiles),¹¹ multivitamin use (yes, no), and fruit and vegetable intake (<3, 3–<5, 5–<10, ≥10 servings/d). When available, we simultaneously adjusted for the reported value of each variable on both the 1982 and 1980 questionnaires.

The Observational Cohort as a Sequence of Nonrandomized Nested “Trials”

The approach described above would produce very imprecise ITT estimates if (as was the case) few women were initiators during the 1982–1984 period. However, our choice of this period was arbitrary. The approach described above can produce an additional NHS nonrandomized “trial” when applied to each of the 8 2-year periods between 1982–1984 and 1996–1998. Thus, as a strategy to increase the efficiency of our ITT estimate, we conducted 7 additional nonrandomized “trials” each subsequent questionnaire (1986, 1988, . . . 1998), and pooled all 8 “trials” into a single analysis. Because some women participated in more than one of these NHS “trials” (up to a maximum of 8), we used a robust variance estimator to account for within-person correlation. We assessed the potential heterogeneity of the ITT effect estimates across “trials” by 2 Wald tests: first, we estimated a separate parameter for therapy initiation in each “trial” and tested for heterogeneity of the parameters (χ^2 ; 6 *df*), and then we calculated a product term (for the indicators of “trial” and therapy initiation), testing for whether the product term was different from 0 (χ^2 ; 1 *df*).

In each “trial,” we used the corresponding questionnaire information to apply the eligibility criteria at the start of follow-up, and to define initiators and noninitiators. We then estimated the CHD average HR in initiators versus noninitiators (adjusted for the values of covariates reported in the 2

previous questionnaires), regardless of whether these women subsequently stopped or initiated therapy. To allow for the possibility that the HR varied with time since baseline, we added product terms between time of follow-up (linear and quadratic terms) and initiation status to a pooled logistic model that approximated our previous Cox model. We then used the fitted model to estimate CHD-free survival curves for initiators and noninitiators.

The subset of women considered for eligibility in each “trial” is approximately nested in the subset of women who were considered for eligibility in the prior “trial.” Our conceptualization of an observational study with a time-varying treatment as a sequence of nested “trials,” each with nontime-varying treatment, is a special case of g-estimation of nested structural models.¹²

Several lines of evidence suggest a modification of the effect of hormone therapy by time of initiation.¹³ We therefore conducted stratified analyses by time since menopause (<10, ≥10 years) and age (<60, ≥60 years). We computed *P* values for “interaction” between hormone therapy and years since menopause by adding a single product term (indicator for hormone therapy times indicator for <10 years since menopause) to the model for the overall HR, and then testing the hypothesis that its coefficient was equal to zero. A less powerful alternative strategy, testing for heterogeneity of the HR estimated from separate models for women <10 years and for women >10 years since menopause, resulted in *P* > 0.15 in all analyses.

Adherence-Adjusted Effect Estimates

Because the primary analysis of the WHI randomized trial was conducted under the ITT principle, we analyzed our NHS “trials” using an observational analog of ITT to compare the NHS with the WHI estimates. However, ITT estimates are problematic because the magnitude of the ITT effect varies with the proportion of subjects who adhere to the assigned treatment, and thus ITT comparisons can underestimate the effect that would have been observed if everyone had adhered to the assigned treatment. Thus, ITT effect estimates may be unsatisfactory when studying the efficacy, and inappropriate when studying the safety, of an active treatment compared with no treatment. An alternative to the ITT effect is the effect that would have been observed if everyone had remained on her initial treatment throughout the follow-up, which we refer to as an adherence-adjusted effect. Under additional assumptions, consistent adherence-adjusted effect estimates can be obtained in both randomized experiments and observational studies by using g-estimation^{14,15} or inverse probability weighting.

We used inverse probability weighting to estimate the adherence-adjusted HR of CHD. In each NHS “trial” we censored women when they discontinued their baseline treatment (either hormone therapy or no hormone therapy), and then weighted the uncensored women months by the inverse

of their estimated probability of remaining uncensored until that month.¹⁶ To estimate “trial”-specific probabilities for each woman, we fit a pooled logistic model for the probability of remaining on the baseline treatment through a given month. The model included the baseline covariates used in the “trial”-specific Cox models described previously, and the most recent postbaseline values of the same covariates. Inclusion of time-dependent covariates is necessary to adjust for any dependence between noncompliance and CHD within levels of baseline covariates. We fit separate models for initiators and noninitiators. In each “trial,” each woman contributed as many observations to the model as the number of months she was on her baseline therapy.

To stabilize the inverse probability weights, we multiplied the weights by the probability of censoring given the trial-specific baseline values of the covariates. Weight stabilization improves precision by helping to reduce random variability. If the true adherence-adjusted HR is constant over time, this method produces valid estimates provided that discontinuing the baseline treatment is unrelated to unmeasured risk factors for CHD incidence within levels of the covariates, and that the logistic model used to estimate the inverse probability weights is correctly specified. When the adherence-adjusted HR changes with time since baseline, this method estimates a weighted average adherence-adjusted HR with time-specific weights proportional to the number of uncensored CHD events occurring at each time. Thus, with heavy censoring due to lack of adherence, the early years of follow-up contribute relatively more weight than would be the case without censoring. To more appropriately adjust for a time-varying HR, we also fit an inverse probability weighted Cox model (approximated through a weighted pooled logistic model) that included product terms between time of follow-up (linear and quadratic terms) and initiation status. We then used the weighted model to estimate adherence-adjusted CHD-free survival curves for initiators and noninitiators.

We also present additional subsidiary analyses to explain the relation between our estimates and previously reported NHS estimates, which can be regarded as estimates of the adherence-adjusted HR using an alternative to our inverse probability weighting approach.

RESULTS

The NHS Nonrandomized “Trials”

Of the 101,819 NHS participants alive and without a history of cancer, heart disease, or stroke in 1984, 81,073 had diet information and, of these, 77,794 were postmenopausal at some time during the follow-up. We excluded 14,764 women who received a form of hormone therapy other than oral estrogen plus progestin in all of the NHS “trials,” or did not provide information on the type of hormone therapy in any of the “trials.” Of the remaining 63,030 women, we excluded 17,146 who received hormone therapy in the 2

years before the baseline of all the “trials.” Of the remaining 45,884 women, we excluded 11,309 who did not have an intact uterus in 1984. Thus 34,575 women met our eligibility criteria for at least one NHS “trial.” Of these women, 1035 had a CHD event, 2596 died of other causes or were lost to follow-up, and 30,944 reached June 2000 free of CHD. Figure 1 shows the distribution of women by number of “trials” in which they participated. Table 1 shows the number of participants, initiators, and CHD events per “trial.” Table 2 shows the distribution of baseline characteristics in initiators and noninitiators.

ITT Estimates of the Effect of Hormone Therapy on CHD

The estimated average HR of CHD for initiators versus noninitiators was 0.96 (95% CI = 0.78–1.18) when the entire follow-up time was included in the analysis (Table 3). The HR was 1.83 (1.05–3.17) when the analysis was restricted to the first year of follow-up, 1.42 (0.92–2.20) for the first 2 years, 1.11 (0.84–1.47) for the first 5 years, and 1.00 (0.78–1.28) for the first 8 years. Equivalently, the HR was 0.96 (0.66–1.39) during years 2–5, 0.81 (0.51–1.28) during years 5–8, and 0.87 (0.58–1.30) after year 8. We did not find a strong indication of heterogeneity across trials (Wald tests *P* values 0.24 and 0.15 for the overall HR). Figure 2A shows that the estimated proportion of women free of CHD during the first 5 years of follow-up was lower in initiators of estrogen plus progestin therapy than in noninitiators of hormone therapy. By year 8, however, this proportion was greater in initiators.

We next examined effect modification, stratifying our ITT estimates by age and time since menopause (Table 3). The HR was 0.84 (CI = 0.61–1.14) in women within 10 years of menopause at baseline, and 1.12 (0.84–1.48) in the others (86% of initiators in this latter group initiated therapy 10 to 20 years after menopause). Similarly, the HRs were 0.86 (0.65–1.14) in women under age 60 at baseline, and 1.15 (0.85–1.57) in the others. Figure 2B, C shows the estimated proportion of women free of CHD by initiator status and time since menopause. The *P* value from a log-rank test for the equality of the survival curves was 0.70 for the entire population, 0.27 for women within 10 years of menopause, and 0.43 for the others.

When we repeated the analyses with no past use of hormone therapy as an additional eligibility criterion (26,797

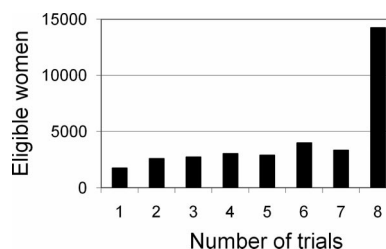


FIGURE 1. Distribution of eligible women by number of Nurses' Health Study “trials” of hormone therapy initiation in which they participated.

TABLE 1. Number of Participants, Therapy Initiators, and CHD Events in Each NHS “Trial” to Estimate the Intention-to-Treat Effect of Initiation of Estrogen/Progestin Therapy

Trial	Questionnaire	Yrs After Menopause	Participants	Initiators	CHD Events	
					Among All Participants	Among Initiators
1	1984	All	16,190	224	772	10
		<10	11,654	201	524	8
		≥10	4536	23	248	2
2	1986	All	17,147	518	671	9
		<10	10,246	416	322	5
		≥10	6901	102	349	4
3	1988	All	18,620	979	610	17
		<10	9550	745	226	8
		≥10	9070	234	384	9
4	1990	All	19,002	1082	528	14
		<10	8014	727	144	5
		≥10	10,988	355	384	9
5	1992	All	19,494	1152	441	13
		<10	7161	714	93	5
		≥10	12,333	438	348	8
6	1994	All	19,954	1344	354	19
		<10	6456	799	69	8
		≥10	13,498	545	285	11
7	1996	All	19,661	1188	228	11
		<10	5508	631	34	3
		≥10	14,153	557	194	8
8	1998	All	18,192	771	100	5
		<10	4287	338	13	2
		≥10	13,905	433	87	3

Based on 34,575 distinct women and 1035 distinct CHD events.

eligible women, 767 CHD events), the HR was 0.79 (CI = 0.60–1.03) for the entire follow-up and 1.49 (0.88–2.54) in the first 2 years (Table 4). The HR was 0.66 (0.44–0.98) in women within 10 years of menopause at baseline, and 1.02 (0.70–1.50) in the others. The appendix includes additional analyses to document the generally small sensitivity of the results regarding the assignment of the month of therapy initiation (Appendix A1), the inclusion of women under age 50 (Appendix A2), the exclusion of women who died between the start of follow-up and the return of the next questionnaire (Appendix A3), the adjustment for confounding by covariates in the proportional hazards model rather than by propensity score methods (Appendix A4), and the assumption of possible unmeasured confounding for therapy discontinuation (Appendix A5).

Adherence-Adjusted Effect Estimates

Figure 3 shows the adherence through year 8 in initiators and noninitiators. The estimated inverse probability weights had mean 1.02 (range = 0.02–30.7) in initiators, and 1.00 (0.17–19.3) in noninitiators. The inverse probability weighted HRs were 0.98 (CI = 0.66–1.49) for the entire follow-up, 1.53 (0.80–2.95) for the first year, 1.61 (0.97–

2.66) for the first 2 years, 1.14 (0.74–1.76) for the first 5 years, and 0.99 (0.66–1.50) for the first 8 years. The HR was 0.65 (0.30–1.38) during years 2 to 5, 0.47 (0.14–1.58), during years 5 to 8, and 0.85 (0.22–3.19) after year 8. The large standard errors that increase with time reflect the fact that few women continued on hormone therapy for long periods. We also examined the effect modification by age and time since menopause (Table 5). Figure 4 shows the estimated adherence-adjusted proportions of women free of CHD. The *P* value from a log-rank test for the equality of the survival curves was 0.91 for the entire population, 0.24 for women within 10 years after menopause, and 0.40 for the others.

Comparison of ITT Estimates With Previous NHS Estimates

The HR estimate of 0.96 from our ITT analysis is not directly comparable with the HR estimate of 0.68 (0.55–0.83) for current users versus never users of estrogen plus progestin reported in the most recent NHS publication.¹⁷ The 0.68 estimate can be interpreted as an adherence-adjusted effect estimate, in which incomplete adherence has been adjusted not by inverse probability weighting but by a comparison of

TABLE 2. Baseline Characteristics of Initiators and Noninitiators of Estrogen/Progestin Therapy in the NHS “Trials”

	Initiators (n = 7258)	Noninitiators (n = 141,002)
Age (y); mean (SD)	59.0 (5.8)	61.9 (6.0)
Age at menopause (y)		
<50	39.2	40.0
50–53	46.6	46.7
>53	14.3	13.4
Years since menopause; mean (SD)	9.0 (5.8)	12.2 (6.3)
<10 y since menopause	63.0	41.4
Past use of hormone therapy	32.0	22.6
Non-Hispanic white	94.6	94.0
Graduate degree	11.7	7.5
Husband's education		
Less than high school	4.8	8.5
High school graduate	36.4	42.8
College	29.9	27.3
Graduate school	29.0	21.4
Parental history of MI before age 60	18.5	17.1
High cholesterol	35.6	32.1
High blood pressure	27.3	33.3
Diabetes	3.5	5.3
Angina	2.1	2.6
Stroke	0.3	0.7
Coronary revascularization	0.7	0.7
Osteoporosis	6.8	6.2
Multivitamin use	44.7	41.8
Cigarette smoking		
Never	41.8	42.0
Past	42.0	36.1
Current	16.2	22.0
Alcohol intake (g/d)		
0	31.7	38.1
0.1–4.9	33.7	31.0
5.0–9.9	12.0	9.9
10.0–14.9	9.9	9.3
≥15	12.7	11.7
Diet score in the 2 upper quintiles	41.6	35.3
Fruit and vegetable intake (servings/d)		
<3	57.9	57.8
3–4.9	29.0	29.0
5–9.9	12.5	12.3
≥10	0.5	0.8
Body mass index (kg/m ²)		
<25	56.2	49.3
25–29.9	29.9	32.1
≥30	13.8	18.7
Aspirin use		
Nonuse	21.9	26.7
1–4 y	18.1	18.9
5–10 y	19.5	15.9
>10 y	40.6	38.5
Physical activity (h/wk)		
<1	43.2	47.1
1–1.9	15.4	15.3
2–3.9	19.8	18.9
4–5.9	10.6	9.4
6–9.9	7.4	6.6
≥10	3.6	3.0

Results are expressed as percentages unless otherwise indicated.

TABLE 3. Estimates of the Intention-to-Treat Effect of Initiation of Estrogen/Progestin Therapy on the Incidence of CHD Events in the NHS “Trials”

	All	Follow-Up Period	
		0–24 Mo	>24 Mo
Initiators			
Total no.	7258	7258	7221
No. CHD events	98	22	76
Noninitiators			
Total no.	141,002	141,002	139,599
No. CHD events	3606	512	3094
	HR (95% CI)	HR (95% CI)	HR (95% CI)
All women	0.96 (0.78–1.18)	1.42 (0.92–2.20)	0.88 (0.69–1.12)
By time after menopause (y)			
<10	0.84 (0.61–1.14)	1.33 (0.66–2.64)	0.77 (0.54–1.09)
≥10	1.12 (0.84–1.48)	1.48 (0.83–2.64)	1.05 (0.77–1.43)
P for interaction	0.08	0.90	0.07
By age (y)			
<60	0.86 (0.65–1.14)	1.36 (0.73–2.52)	0.78 (0.57–1.07)
≥60	1.15 (0.85–1.57)	1.49 (0.79–2.80)	1.08 (0.76–1.54)
P for interaction	0.05	0.72	0.06

Adjusted for the following baseline variables: age, parental history of myocardial infarction before age 60, education, husband's education, ethnicity, age at menopause, calendar month, high cholesterol, high blood pressure, diabetes, angina, stroke, coronary revascularization, osteoporosis, body mass index, cigarette smoking, aspirin use, alcohol intake, physical activity, diet score, multivitamin use, fruits and vegetables intake, and previous use of hormone therapy. The last column is restricted to women who were still at risk after the first 2 years of follow-up of the corresponding trial.

current versus never users. This approach is used in many large observational cohorts, including the NHS (see “Discussion” for details). Table 6 shows the cumulative steps that link our estimates in Table 3 with the previously reported NHS estimate. These steps involve changes in the start of follow-up, the definition of the exposed and unexposed group, the covariates used for adjustment, and eligibility criteria.

Column i of Table 6 shows the estimates when (as in previous NHS analyses) the start of follow-up, and thus the “baseline,” of each trial was redefined as the date of return of the questionnaire. When “baseline” is modified in this way, the selected group of initiators differs from the initiator group in Table 3 because it does not include women who, during the 2-year interval before “baseline,” either initiated and stopped hormone therapy or survived a CHD event occurring after initiation. As in Table 3, we provide separate HR estimates for the entire follow-up (0.84), the first 2 years of follow-up (0.98), and the period after the first 2 years (0.80).

Second, we varied the definition of the user and non-user groups in 3 steps as shown in the next 3 columns of Table 6. In column ii we eliminated our “trial”-specific criterion of no therapy in the 2 years before “baseline” for initiators; that is, we compared current users with noninitia-

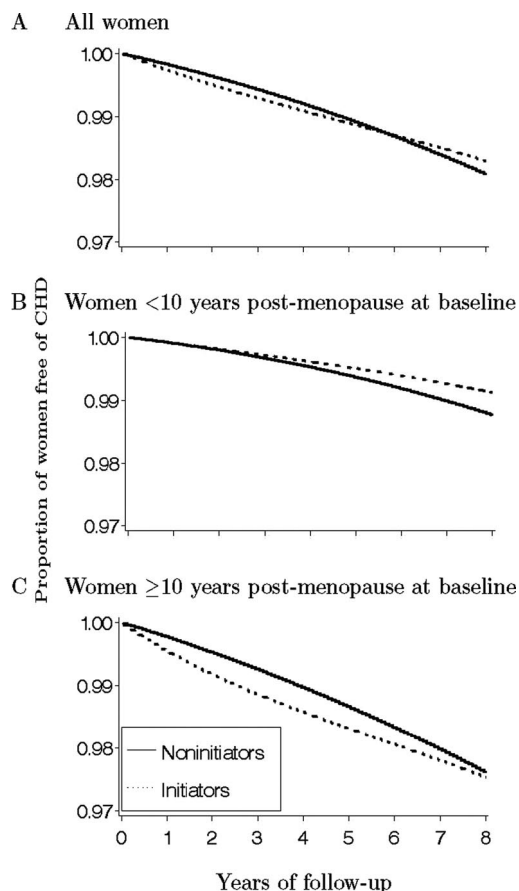


FIGURE 2. Proportion of women free of CHD by baseline treatment group in the Nurses' Health Study "trials."

tors. In column iii we eliminated our "trial"-specific criterion of no therapy in the 2 years before "baseline" for all women; that is, we compared current users with current nonusers. In column iv we used as the comparison group the subset of nonusers with no history of hormone therapy use; that is, we compared current users with never users as in previous NHS analyses. The HR estimates for columns ii, iii, iv were, respectively, 0.84, 0.86, 0.85 for the entire follow-up, 0.77, 0.77, 0.74 for 0 to 24 months, and 0.87, 0.90, 0.90 for >24 months.

To explain why the number of exposed cases ($n = 319$) in columns ii to iv far exceeds the number ($n = 66$) in column i, consider a woman who is continuously on hormone therapy from 1982–1984 until she dies of CHD just before the end of follow-up in 2000. In the analysis of column i, this woman participates as an exposed CHD case in the first (1984) "trial" only. In contrast, in the analyses of columns ii to iv, the same woman participates as an exposed CHD case in each of the 8 "trials" 1984–1998. Furthermore, in the analysis of column i, the woman would contribute 0 to the 0- to 24-month exposed case stratum and 1 to the >24-month exposed case stratum. In contrast, the same woman in the analyses of columns ii to

TABLE 4. Estimates of the Intention-to-Treat Effect of Initiation of Estrogen/Progestin Therapy on the Incidence of CHD Events Among Women With No History of Hormone Use in the NHS "Trials"

		Follow-Up Period	
	All	0–24 Mo	>24 Mo
Initiators			
Total no.	4939	4939	4914
No. CHD events	55	15	40
Noninitiators			
Total	109,205	109,205	108,108
No. CHD events	2723	379	2344
	HR (95% CI)	HR (95% CI)	HR (95% CI)
All women	0.79 (0.60–1.03)	1.49 (0.88–2.54)	0.68 (0.49–0.93)
By time after menopause (y)			
<10	0.66 (0.44–0.98)	1.32 (0.58–3.03)	0.58 (0.37–0.90)
≥10	1.02 (0.70–1.50)	1.71 (0.85–3.45)	0.88 (0.56–1.38)
<i>P</i> for interaction	0.09	0.51	0.15
By age (y)			
<60	0.68 (0.48–0.97)	1.38 (0.66–2.88)	0.59 (0.39–0.88)
≥60	1.06 (0.69–1.64)	1.64 (0.73–3.69)	0.93 (0.56–1.56)
<i>P</i> for interaction	0.10	0.69	0.13
Adjusted for same baseline variables as in Table 3.			

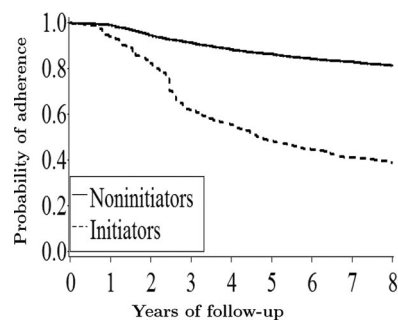


FIGURE 3. Proportion of women who adhered to their baseline treatment in the Nurses' Health Study "trials."

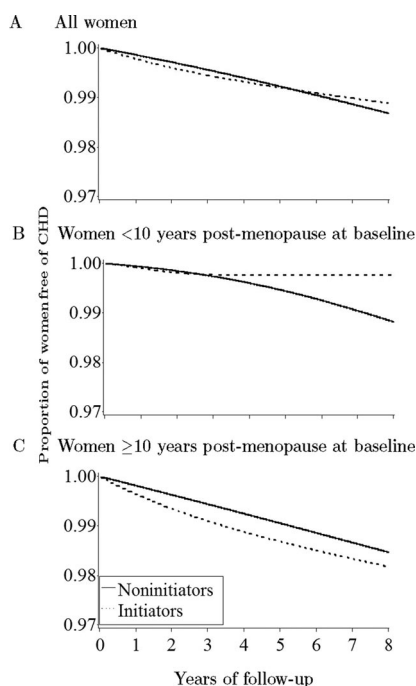
iv would contribute 1 to the 0- to 24-month exposed case stratum (corresponding to the 1998 "trial") and 7 to the >24-month exposed case stratum (corresponding to each of the other 7 "trials").

Third, we repeated the analysis in column iv after adjusting for the set of covariate values used in the most recent NHS publication. Thus, the estimates in column v—0.81 for the entire follow-up, 0.71 for 0 to 24 months, and 0.85 for >24 months—were adjusted for the most recent values available at the time of return of the "baseline" questionnaire, rather than the most recent values available at the 2 previous questionnaires.

TABLE 5. Estimates of the (Adherence-Adjusted) Effect of Continuous Estrogen/Progestin Therapy Versus No Hormone Therapy on the Incidence of CHD Events in the NHS “Trials”

	All HR (95% CI)	Follow-Up Period	
		0–24 Mo HR (95% CI)	>24 Mo HR (95% CI)
All women	0.98 (0.66–1.45)	1.61 (0.97–2.66)	0.64 (0.35–1.15)
By time after menopause (y)			
<10	0.54 (0.19–1.51)	1.21 (0.40–3.61)	0.14 (0.02–1.16)
≥10	1.20 (0.78–1.84)	1.92 (1.09–3.39)	0.84 (0.45–1.56)
<i>P</i> for interaction	0.01	0.18	0.11
By age (y)			
<60	0.78 (0.44–1.40)	1.65 (0.81–3.37)	0.45 (0.19–1.09)
≥60	1.36 (0.81–2.29)	1.69 (0.86–3.32)	1.08 (0.50–2.36)
<i>P</i> for interaction	0.06	0.74	0.09

Adjusted for same baseline variables as in Table 3. In each “trial,” women were censored when they discontinued their baseline treatment (either hormone therapy or no hormone therapy), and the uncensored women months were weighted by the inverse of their estimated probability of remaining uncensored until that month.

**FIGURE 4.** Proportion of women free of CHD under full adherence with the baseline treatment in the Nurses' Health Study “trials.”

Fourth, we repeated the analysis in column v after dropping the requirement of an intact uterus, which was not used in previous NHS analyses. The estimates in column vi were 0.82 for the entire follow-up, 0.67 for 0 to 24 months, and 0.87 for >24 months. The estimate 0.67 in the row 0 to 24 months corresponds almost exactly to the analytic approach used in the most recent NHS publication,¹⁷ which

estimated the HR over the 2-year period after the reclassification (ie, updating) of treatment status at the return of each questionnaire.

DISCUSSION

We used the NHS observational data to emulate the design and analysis of the WHI randomized trial. The ITT HRs of CHD for therapy initiation were 1.42 (95% CI = 0.92–2.20) in the NHS vs. 1.68 (95% CI = 1.15–2.45) in the WHI⁹ during the first 2 years, and 1.00 (0.78–1.28) in the NHS versus approximately 1.24 (0.97–1.60) in the WHI⁸ during the first 8 years. However, much of the apparent WHI-NHS difference disappeared after stratification by time since menopause at hormone therapy initiation. The ITT HRs were 0.84 (0.61–1.14) in the NHS versus 0.88 (0.54–1.43) in the WHI^{8,18} for women within 10 years after menopause, and approximately 1.12 (0.84–1.48) in the NHS versus 1.23 (0.85–1.77) in the WHI^{8,18} for women between 10 and 20 years after menopause.

These findings provide additional support to the hypothesis that hormone therapy may increase the long-term CHD risk only in women who were 10 or more years after menopause at initiation,^{17,19} and to the rationale for an ongoing randomized clinical trial to determine the effect of estrogen plus progestin on coronary calcification in younger women.²⁰ When the analyses were limited to women with no history of hormone use, the ITT HR was 0.79 (0.60–1.03) for the entire follow-up and 0.66 (0.44–0.98) for women who initiated hormone use within 10 years of menopause.

We computed average ITT HRs in the NHS for comparison with the main result of the WHI. Our ITT estimates suggest that any remaining differences between NHS and WHI estimates are not explained by unmeasured joint risk factors for CHD and therapy discontinuation. However, the average ITT HR is not the ideal effect measure because the survival curves crossed during the follow-up in both the WHI trial and the NHS trials, and also because ITT estimates like the ones shown here are generally attenuated toward the null due to misclassification of actual treatment. We addressed the first problem by estimating survival curves to first CHD event, and the second problem by estimating these curves under full adherence (via inverse probability weighting). Therefore the adherence-adjusted survival curves of Figure 4 provide the most appropriate summary of our results. It will be of interest to compare these results with adherence-adjusted curves (via inverse probability weighting) from the WHI when they become available. The curves suggest that continuous hormone therapy causes a net reduction in CHD among women starting therapy within 10 years of menopause, and a net increase among those starting later. However, either of these effects could be due to sampling variability.

Previously published NHS estimates¹⁷ compared the hazards of current versus never users over the 2-year period

TABLE 6. Comparison of Several Alternative Hazard Ratio Estimates With the Previously Reported Estimate From the NHS (Column vi, Row 0–24 Mo)

					Current Users vs. Never Users		
	Initiators vs. Noninitiators ^a	Selected ^b Initiators vs. Noninitiators (i)	Current Users vs. Noninitiators (ii)	Current Users vs. Nonusers (iii)	(iv)	Covariates of Previous NHS Analyses (v)	Not Requiring Presence of Uterus (vi)
Entire Follow-up							
Users							
Total no.	7258	6400	41,441	41,441	41,441	41,441	45,793
No. CHD events	98	66	319	319	319	319	398
Nonusers							
Total no.	141,002	141,316	141,316	173,094	126,235	126,235	147,045
No. CHD events	3606	3271	3271	3764	2778	2778	3404
All women	0.96 (0.78–1.18)	0.84 (0.64–1.09)	0.84 (0.67–1.06)	0.86 (0.70–1.06)	0.85 (0.68–1.07)	0.81 (0.65–1.01)	0.82 (0.68–0.99)
Time from menopause							
<10 y	0.84 (0.61–1.14)	0.66 (0.45–0.98)	0.76 (0.57–1.02)	0.79 (0.60–1.03)	0.76 (0.57–1.01)	0.74 (0.56–0.98)	0.77 (0.59–1.00)
≥10 y	1.12 (0.84–1.48)	1.05 (0.75–1.47)	0.95 (0.72–1.27)	0.95 (0.72–1.25)	0.92 (0.68–1.25)	0.90 (0.68–1.20)	0.89 (0.70–1.13)
P interaction	0.08	0.03	0.05	0.09	0.08	0.08	0.12
Age							
<60 y	0.86 (0.61–1.14)	0.67 (0.47–0.97)	0.80 (0.61–1.05)	0.82 (0.64–1.06)	0.79 (0.60–1.03)	0.79 (0.60–1.05)	0.79 (0.62–1.01)
≥60 y	1.15 (0.85–1.57)	1.14 (0.80–1.63)	0.92 (0.67–1.26)	0.94 (0.69–1.27)	0.93 (0.67–1.29)	0.83 (0.63–1.10)	0.86 (0.68–1.10)
P interaction	0.05	0.01	0.14	0.20	0.17	0.49	0.24
0–24 mo							
CHD events							
No. users	22	17	80	80	80	80	90
No. nonusers	512	660	660	755	542	542	666
HR (95% CI)	1.42 (0.92–2.20)	0.98 (0.60–1.60)	0.77 (0.60–0.99)	0.77 (0.60–0.98)	0.74 (0.57–0.95)	0.71 (0.55–0.91)	0.67 (0.54–0.85)
>24 mo							
CHD events							
No. users	76	49	239	239	239	239	308
No. nonusers	3094	2611	2611	3008	2236	2236	2738
HR (95% CI)	0.88 (0.69–1.12)	0.80 (0.60–1.08)	0.87 (0.68–1.12)	0.90 (0.72–1.14)	0.90 (0.70–1.15)	0.85 (0.67–1.08)	0.87 (0.71–1.07)

^aFrom Table 3. Follow-up starts at time of therapy initiation. In all other columns follow starts at time of questionnaire return.

^bWomen who initiated and stopped therapy, or who survived a CHD event, between the time of therapy initiation and the time of questionnaire return are excluded. See main text for a description of each estimate.

after the updating of treatment status at the return of each questionnaire, and could thus be viewed as a form of adherence adjustment. In Table 6 we described the steps from our 2-year ITT estimate to the previously published adherence-adjusted estimate. Below we discuss the 2 key steps: the change of start of follow-up (time of therapy initiation vs. time of questionnaire return), and the change of the exposed group (selected initiators vs. current users).

The 2-year HR estimate changed from 1.42 (Table 3) to 0.98 (Table 6, column i) during the first 2 years, and from 0.96 (Table 3) to 0.84 (Table 6, column i) for the entire follow-up when the definition of start of follow-up was changed from the estimated time of therapy initiation to the time of return of the next questionnaire (the latter definition is commonly used in observational studies that collect treatment information at regular intervals). This latter definition excludes women who initiated treatment and then suffered a nonfatal myocardial infarction during the interval between

treatment initiation and treatment ascertainment (up to 2 years in the NHS). If hormone therapy increases the short-term risk of CHD, this exclusion will result in an underestimation of the early increase in risk and may result in selection bias,¹⁶ which may explain part of the change from 1.42 to 0.98. The impact of this exclusion bias, however, will be diluted over the entire follow-up, as previously suggested in a sensitivity analysis,¹⁷ which may explain the smaller change from 0.96 to 0.84. This exclusion bias may be quantified through simulations,²¹ reduced by stratification of the analysis on duration of therapy at baseline,²¹ and eliminated by making the start of follow-up coincident with the time of treatment initiation, as discussed by Robins^{22,23} and Ray.²⁴ The approach we present here and elsewhere^{10,25} generalizes Ray's "new-users design" to the case of time-varying treatments.

The point estimate further changed from 0.98 (Table 6, column i) to 0.77 (column ii) when the definition of exposure

changed from selected initiators to current users. These are estimates for different contrasts. The estimate in column i is based on the exposed person-time during the 2-year period immediately after the return of the questionnaire in which therapy initiation was reported, and thus can be viewed as a flawed attempt to estimate the early effect of therapy initiation (see previous paragraph). The estimate in column ii, however, is based on the exposed person-time pooled over all 2-year periods after the return of any questionnaire, and thus can be interpreted as an attempt to estimate the effect of therapy use during any 2-year period (that excludes the interval between therapy initiation and return of the next questionnaire, as discussed in the previous paragraph). More specifically, the approach in column ii can be understood as an attempt to estimate adherence-adjusted effects by entering the current value of exposure and the joint predictors of adherence and CHD as time-varying covariates in the model for CHD risk. Unlike inverse probability weighting, this approach to adherence adjustment requires that the time-dependent covariates not be strongly affected by prior treatment. This may be a reasonable assumption in the NHS. Thus the estimates in column ii may be more usefully compared with a weighted average of our interval-specific adherence adjusted estimates of 1.61 (0–2 years), 0.65 (2–5 years), 0.47 (5–8 years), and 0.85 (>8 years) than to the estimate in column i.

In summary, our findings suggest that the discrepancies between the WHI and NHS ITT estimates could be largely explained by differences in the distribution of time since menopause and length of follow-up. Residual confounding for the effect of therapy initiation in the NHS seems to play little role.

ACKNOWLEDGMENTS

We thank Murray Mittleman, Javier Nieto, Meir Stampfer, and Alexander Walker for their comments on an earlier version of the manuscript.

REFERENCES

- Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*. 2001;286:821–830.
- Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000;342:1878–1886.
- Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342:1887–1892.
- Grodstein F, Stampfer M, Manson J, et al. Postmenopausal estrogen and progestin use and the risk of cardiovascular disease [Erratum in: *N Engl J Med*. 1996;335:1406]. *N Engl J Med*. 1996;335:453–461.
- Grodstein F, Manson JE, Colditz GA, et al. A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. *Ann Intern Med*. 2000;133:933–941.
- Varas-Lorenzo C, García-Rodríguez LA, Pérez-Gutthann S, et al. Hormone replacement therapy and incidence of acute myocardial infarction. *Circulation*. 2000;101:2572–2578.
- The Women's Health Initiative Study Group. Design of the women's health initiative clinical trial and observational study. *Control Clin Trials*. 1998;19:61–109.
- Manson JE, Hsia J, Johnson KC, et al. Estrogen plus progestin and the risk of coronary heart disease. *N Engl J Med*. 2003;349:523–534.
- Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. *Biometrics*. 2005;61:899–911.
- Hernán MA, Robins JM, García Rodríguez LA. In discussion of: Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. *Biometrics*. 2005;61:922–930.
- Stampfer MJ, Hu FB, Manson JE, et al. Primary prevention of coronary heart disease in women through diet and lifestyle. *N Engl J Med*. 2000;343:16–22.
- Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Services Research Methodology: A Focus on AIDS: NCHRS, US*. Washington, DC: Public Health Service;1989:113–159.
- Mendelsohn ME, Karas RH. Hormone replacement therapy and the young at heart. *N Engl J Med*. 2007;356:2639–2641.
- Mark SD, Robins JM. A method for the analysis of randomized trials with compliance information: an application to the Multiple Risk Factor Intervention Trial. *Control Clin Trials*. 1993;14:79–97.
- Cole SR, Chu H. Effect of acyclovir on herpetic ocular recurrence using a structural nested model. *Contemp Clin Trials*. 2005;26:300–310.
- Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
- Grodstein F, Manson JE, Stampfer MJ. Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. *J Women's Health*. 2006;15:35–44.
- Manson JE, Bassuk SS. Invited commentary: hormone therapy and risk of coronary heart disease why renew the focus on the early years of menopause? *Am J Epidemiol*. 2007;166:511–517.
- Grodstein F, Clarkson TB, Manson JE. Understanding the divergent data on postmenopausal hormone therapy. *N Engl J Med*. 2003;348:645–650.
- Harman SM, Brinton EA, Cedars M, et al. KEEPS: The Kronos Early Estrogen Prevention Study. *Climacteric*. 2005;8:3–12.
- Prentice RL, Langer RD, Stefanick ML, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the women's health initiative clinical trial. *Am J Epidemiol*. 2005;162:404–414.
- Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period — Application to the healthy worker survivor effect [published errata appear in *Math Model*. 1987;14:917–921]. *Math Model*. 1986;7:1393–1512.
- Robins JM. Addendum to “A new approach to causal inference in mortality studies with a sustained exposure period—application to the healthy worker survivor effect” [published errata appear in *Comput Math Appl*. 1989;18:477]. *Comput Math Appl*. 1987;14:923–945.
- Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158:915–920.
- Alonso A, García Rodríguez LA, Logroscino G, et al. Gout and risk of Parkinson's disease: a prospective study. *Neurology*. 2007;69:1696–1700.
- Connelly M, Richardson M, Platt R. Prevalence and duration of postmenopausal hormone replacement therapy use in a managed care organization. *J Gen Intern Med*. 2000;15:542–50.
- Robins JM, Rotnitzky A, Vansteelandt S. In discussion of: Frangakis CE, Rubin DB, An M, MacKenzie E. “Principal stratification designs to estimate input data missing due to death”. *Biometrics*. 2007;63:650–662.
- Robins JM. Structural nested failure time models. In: Armitage P, Colton T, eds. *Survival Analysis. The Encyclopedia of Biostatistics*. Chichester, UK: John Wiley and Sons; 1998:4372–4389.
- Hernán MA, Cole SR, Margolick J, et al. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol Drug Saf*. 2005;14:477–491.
- Robins JM, Blevins D, Ritter G, et al. G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients [published errata appear in *Epidemiology*. 1993;4:189]. *Epidemiology*. 1992;3:319–336.

APPENDIX: SENSITIVITY TO OUR ANALYTIC CHOICES FOR THE NHS NONRANDOMIZED TRIALS

We now describe the estimates from sensitivity analyses that alter some of the decisions we made for the analyses shown in Table 3. The results from these sensitivity analyses indicate that these decisions had only a moderate influence on our estimates.

Appendix A1: The Determination of Month of Therapy Initiation

The duration of use of hormone therapy during a given 2-year period is ascertained as a categorical variable with 5 levels in the NHS questionnaires. Therefore any decisions regarding the exact month of therapy initiation will result in some error. We explored the sensitivity of our estimates to this error by conducting separate analyses in which we varied the decisions used to obtain the estimates in Table 3. In the analyses shown in Appendix Table 1, we used the latest possible month of initiation as the month of therapy initiation. For example, if a woman on hormone therapy reported 15–19 months of use during the 2-year period before the return of the baseline questionnaire, we calculated the month of initiation as the month of questionnaire return minus 19 in Table 3, and minus 15 in Appendix Table 1.

APPENDIX TABLE 1. Estimates of the Intention-to-Treat Effect of Initiation of Estrogen/Progestin Therapy on the Incidence of CHD Events When Using the Latest Possible Month of Therapy Initiation in the NHS “Trials”

		Follow-up Period	
	All	0–24 Mo	>24 Mo
Initiators			
Total no.	7245	7245	7165
No. CHD events	90	24	66
Noninitiators			
Total no.	140,881	140,881	139,331
No. CHD events	3533	545	2988
	HR (95% CI)	HR (95% CI)	HR (95% CI)
All women	0.92 (0.74–1.15)	1.49 (0.97–2.27)	0.81 (0.63–1.05)
By time after menopause (y)			
<10	0.81 (0.59–1.12)	0.99 (0.44–2.20)	0.79 (0.55–1.12)
≥10	1.06 (0.79–1.43)	1.84 (1.11–3.05)	0.88 (0.62–1.26)
P for interaction	0.11	0.20	0.35
By age (y)			
<60	0.81 (0.60–1.09)	1.04 (0.51–2.10)	0.76 (0.55–1.07)
≥60	1.13 (0.82–1.56)	1.98 (1.16–3.40)	0.93 (0.63–1.38)
P for interaction	0.04	0.11	0.22

Adjusted for same baseline variables as in Table 3.

Appendix A2: The Inclusion of Women Over Age 50

The WHI trial excluded women younger than 50 years at baseline. Appendix Tables 2 and 3 show, respectively, the ITT and adherence-adjusted estimates when we added this exclusion criterion to the eligibility criteria of our NHS “trials.” The ITT HRs (95% CIs) of CHD for initiators versus non-initiators were 0.99 (0.80–1.22) for the entire follow-up, 1.80 (1.01–3.19) for the first year, 1.43 (0.92–2.23) for the first 2 years, 1.13 (0.85–1.50) for the first 5 years, and 1.05 (0.82–1.34) for the first 8 years. The adherence-adjusted HRs (95% CIs) were 1.30 (0.76–2.21) for the entire follow-up, 1.61 (0.84–3.08) for the first year, 1.71 (1.03–2.83) for the first 2 years, 1.22 (0.80–1.88) for the first 5 years, and 1.35 (0.78–2.35) for the first 8 years. The HR (95% CI) was 0.69 (0.32–1.48) during years 2–5, 1.73 (0.41–2.11) during years 5–8, and 0.91 (0.17–4.83) after year 8.

Appendix A3: The Exclusion of Women Who Died Between the Start of Follow-Up and the Return of the Next Questionnaire

There are 2 reasons why the initiators in our analysis were actually a selected group of all initiators. First, it is possible that some short-term users of hormone therapy were not detected in the NHS. Of note, the adherence of NHS women during the first year after initiation was higher than that previously found in other US²⁶ and UK¹⁰ women, which might reflect a truly greater adherence of NHS women or the

APPENDIX TABLE 2. Estimates of the Intention-to-Treat Effect of Initiation of Estrogen/Progestin Therapy on the Incidence of CHD Events Among Women Aged 50 or More at Baseline in the NHS “Trials”

		Follow-up Period	
	All	0–24 Mo	>24 Mo
Initiators			
Total no.	6602	6602	6566
No. CHD events	94	21	73
Noninitiators			
Total no.	135,877	135,887	134,491
No. CHD events	3503	503	3000
	HR (95% CI)	HR (95% CI)	HR (95% CI)
All women	0.99 (0.80–1.22)	1.43 (0.92–2.23)	0.91 (0.72–1.16)
By time after menopause (y)			
<10	0.88 (0.63–1.21)	1.28 (0.62–2.64)	0.81 (0.56–1.17)
≥10	1.13 (0.85–1.49)	1.50 (0.84–2.68)	1.06 (0.77–1.44)
P for interaction	0.12	0.85	0.11
By age (y)			
<60	0.89 (0.67–1.19)	1.36 (0.71–2.57)	0.82 (0.59–1.14)
≥60	1.15 (0.85–1.57)	1.49 (0.79–2.80)	1.08 (0.77–1.53)
P for interaction	0.08	0.73	0.09

Adjusted for same baseline variables as in Table 3.

APPENDIX TABLE 3. Estimates of the (Adherence-Adjusted) Effect of Continuous Estrogen/Progestin Therapy Versus No Hormone Therapy on the Incidence of CHD Events Among Women Aged 50 or More at Baseline in the NHS "Trials"

	All HR (95% CI)	0–24 Mo HR (95% CI)	>24 Mo HR (95% CI)
All women	1.30 (0.76–2.21)	1.71 (1.03–2.83)	1.07 (0.44–2.63)
By time after menopause (y)			
<10 y	0.68 (0.24–1.91)	1.28 (0.43–3.86)	0.20 (0.03–1.54)
≥10 y	1.57 (0.86–2.85)	1.97 (1.11–3.47)	1.37 (0.54–3.45)
<i>P</i> for interaction	0.03	0.37	0.06
By age (y)			
<60	0.91 (0.49–1.69)	1.80 (0.83–3.87)	0.54 (0.20–1.49)
≥60	1.92 (0.90–4.10)	1.69 (0.87–3.32)	2.10 (0.68–6.50)
<i>P</i> for interaction	0.06	0.94	0.09

Adjusted for same baseline variables as in Table 3. In each "trial," women were censored when they discontinued their baseline treatment (either hormone therapy or no hormone therapy), and the uncensored women-months were weighted by the inverse of their estimated probability of remaining uncensored until that month.

questionnaires' inability to identify all short-term users. Second, both the initiators (and noninitiators) in our analysis did not include women who died before returning the questionnaire. The month of therapy initiation, if any, for women who died between the start of follow-up and the return of the next questionnaire is unknown. As a result, these women were not included in our analyses in Table 3, which might have resulted in selection bias if the women who had a CHD event and died before returning the questionnaire were more (or less) likely to have initiated therapy than those who did not die. As an aside, because the analyses presented in columns i–vi of Table 6 used the date of return of the questionnaire as the start of follow-up, the number of women excluded for this reason is lower in Table 6 than in Table 3. This explains why the number of CHD cases during the first 2 years of follow-up is 534 in Table 3 and 677 in column i of Table 6.

We used inverse probability weighting¹⁶ to adjust for the potential selection bias due to death before questionnaire return. Specifically, we estimated the conditional probability of surviving until the return of the questionnaire for every woman who, having had a CHD event during the 2-year interval prior to the baseline questionnaire, survived to return the questionnaire. We then upweighted these survivors by the inverse of their estimated conditional probability of survival. This approach implicitly assumes that there exists a hypothetical intervention to prevent death before returning the questionnaire among women who had a CHD event.

To estimate the probability of survival, we fit a logistic model among women who had a CHD event in the 2-year interval before the return of the questionnaire. The outcome of the model was the probability of survival until questionnaire return, and the covariates were those used in our Table 3 analyses to adjust for confounding. This approach adjusts only for the selection bias that can be explained by these

covariates. Appendix Table 4 shows the inverse probability weighted ITT HRs and their 95% CIs, which are similar to those in Table 3—although the HR for initiators versus noninitiators during the first 2 years of follow-up was closer to the null in Appendix Table 4 (1.30) than in Table 3 (1.48).

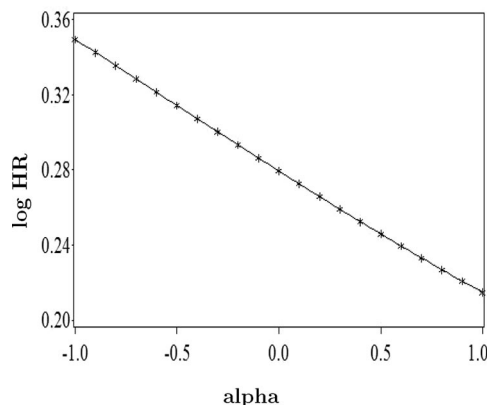
However, our inverse probability weighted analysis could not adjust for treatment status because it is unknown whether women who died before returning their questionnaire were initiators. Thus, if the probability of dying after or from a CHD event was affected by treatment, our inverse probability weighted analysis would not appropriately adjust for the selection bias. We conducted a sensitivity analysis to determine whether lack of adjustment for treatment status could explain the increased CHD incidence observed in initiators during the first 2 years of follow-up. The methodology for this sensitivity analysis has been recently described.²⁷ Appendix Figure 1 summarizes the results.

The ITT HR of CHD varies from 1.42 for $\alpha = -1$ to 1.24 for $\alpha = 1$, where α is the log odds ratio for the hypothesized association between treatment arm and death before returning the questionnaire, conditional on the other covariates. Our analysis in Appendix Table 4 corresponds to $\alpha = 0$. These results suggest that the potential selection bias due to lack of adjustment for treatment arm in the inverse probability-weighted analysis does not fully explain the increased CHD incidence rate during the first 2 years of follow-up in initiators versus noninitiators.

APPENDIX TABLE 4. Estimates of the Intention-to-Treat Effect of Initiation of Estrogen/Progestin Therapy on the Incidence of CHD Events in the NHS "Trials," After Adjustment for Exclusion of Women Who Died Between the Start of Follow-Up and the Return of the Next Questionnaire

	All	Follow-up Period	
		0–24 Mo	>24 Mo
Initiators			
Total no.	7258	7258	7221
No. CHD events	98	22	76
Noninitiators			
Total no.	141,002	141,002	139,599
No. CHD events	3606	512	3094
	HR (95% CI)	HR (95% CI)	HR (95% CI)
All women	0.96 (0.77–1.18)	1.30 (0.83–2.05)	0.88 (0.69–1.12)
By time after menopause (y)			
<10	0.85 (0.62–1.16)	1.37 (0.69–2.73)	0.77 (0.54–1.09)
≥10	1.09 (0.82–1.46)	1.27 (0.69–2.34)	1.05 (0.77–1.43)
<i>P</i> for interaction	0.12	0.82	0.07
By age (y)			
< 60	0.86 (0.65–1.14)	1.33 (0.71–2.47)	0.78 (0.57–1.07)
≥60	1.14 (0.83–1.55)	1.31 (0.68–2.53)	1.08 (0.76–1.54)
<i>P</i> for interaction	0.07	0.93	0.06

Adjustment does not affect to estimates for >24 mo. HR adjusted for same baseline variables as in Table 3.



APPENDIX FIGURE 1. Sensitivity analysis for lack of adjustment for treatment arm in the inverse probability weighted analysis that adjusts for selection bias due to death between the start of follow-up and the return of questionnaire in the Nurses' Health Study "trials." The parameter alpha is the log odds ratio for the hypothesized association between treatment arm and death before returning the questionnaire. Log HR is the log HR of CHD for initiators versus noninitiators during the first 2 years of follow-up.

Appendix A4: The Use of Propensity Scores

To assess whether our results were affected by the choice of the effect measure (ie, HR) or by the method of adjustment for confounding, we also conducted the analyses by g-estimation of a nested, trial-specific, time-independent accelerated failure time model,^{10,28} which estimates the median survival time ratio of noninitiators versus initiators and adjusts for confounding by combining a model for the propensity score with a model for the effect of the covariates on time to CHD.²⁹ G-estimation of nested structural models is a particularly robust way of utilizing propensity scores as it is minimally affected by poor overlap in the propensity scores of the treated and untreated.^{29,30} The estimates, shown in Appendix Table 5, are qualitatively similar to those in Table 3, which suggests that our conclusions are not sensitive to the method used for confounding adjustment.

Appendix A5: The Assumption of No Unmeasured Confounding

To examine the amount of confounding by measured lifestyle and socioeconomic compared with other risk factors, we first repeated the analysis in Table 3 without adjusting for measured lifestyle factors (alcohol intake, physical activity, aspirin use, diet score, multivitamin use, fruit and vegetable intake). The HR was 0.94 (95% CI = 0.76–1.16). When we also omitted adjustment for our measures of socioeconomic status (education, ethnicity, husband's education), the HR was 0.92 (0.75–1.14). We repeated the analyses without adjusting for any of the potential confounders except age; the age-adjusted HR was 0.67 (0.54–0.83) for CHD. Finer stratification by age (in 2-year intervals) and adjustment for age as a continuous covariate did not materially affect the results.

APPENDIX TABLE 5. Estimates of the Intention-to-Treat Effect of Initiation of Estrogen/Progestin Therapy on the Incidence of CHD Events in the NHS "Trials" (Effect Measured as Median Survival Time Ratio)

Entire Follow-up	
Initiators	
Total no.	7258
No. CHD events	98
Noninitiators	
Total no.	141,002
No. CHD events	3,606
STR (95% CI)	0.87 (0.62–1.08)
By time after menopause (y)	
<10	0.71 (0.43–1.03)
≥10	1.04 (0.70–1.36)
By age (y)	
<60 y	0.66 (0.47–1.03)
≥60 y	1.11 (0.72–1.50)
0–24 mo	
CHD events no.	
Initiators	22
Noninitiators	512
STR (95% CI)	1.82 (0.50–3.70)
Survival time ratios (STRs) adjusted for same baseline variables as in Table 3.	

It is suspected that important confounders of the effect of hormone therapy on CHD risk also confound its effect on stroke risk. Thus we estimated the ITT effect of hormone therapy on stroke under the hypothesis that, in the presence of substantial unmeasured confounding for the effect on CHD risk, the effect estimates for stroke would also be biased. There were 574 cases of stroke among eligible women. Applying the same analytic strategy as in Table 3, the overall HR for stroke was 1.39 (CI = 1.09–1.77), which is similar to the estimate found in the WHI randomized trial.

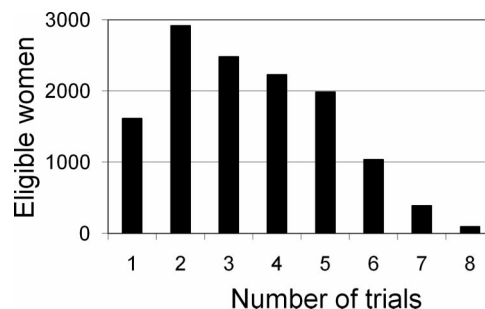
We also repeated the analysis in column vi of Table 6 without adjustment for measured lifestyle factors other than smoking (alcohol intake, physical activity, aspirin use, multivitamin use, vitamin E intake). The HR was 0.67 (CI = 0.53–0.85). When we also omitted adjustment for our measures of socioeconomic status (husband's education), the HR was 0.65 (0.52–0.82). We repeated the analyses without adjusting for any of the potential confounders except age; the age-adjusted HR was 0.48 (0.38–0.60).

To further evaluate whether our decision not to assume comparability on unmeasured factors between those continuing versus discontinuing therapy had an important effect on our adherence-adjusted estimates, we compared our estimated ITT effect of hormone initiation with an estimate of the ITT effect of discontinuation under the assumption of no unmeasured confounders for discontinuation. To calculate this latter effect we recreated a set of NHS "trials" with the same protocol and analytic approach described above except that we restricted participation in each "trial" to women who

reported use of hormone therapy in the questionnaire before baseline.

We implemented the ITT approach by considering the treatment variable to be either 1 or 0 depending on whether the woman reported herself to be off versus on hormone therapy at the baseline questionnaire (regardless of future hormone history), and fit the Cox models described above. Under the assumption of no unmeasured confounders for treatment discontinuation given the variables used in our analysis, the estimates of effect so obtained are comparable with those from a randomized trial among hormone users in which treatment discontinuation is assigned at random.

Our analyses included 12,739 women who met the eligibility criteria for at least 1 NHS estrogen/progestin discontinuation “trial.” Appendix Figure 2 shows the distribution of women by number of “trials” in which they participated. Of these, 131 had a CHD event, 49 died of other causes or were lost to follow-up, and 12,559 reached the administrative end of follow-up free of a diagnosis of CHD. Appendix Table 6 shows the number of participants, stoppers, and CHD events in each of the “trials,” which include fewer participants than those for hormone therapy initiation because they are restricted to the smaller group of hormone therapy users. The HR when we compared the 52 events in the 4617 stoppers with the 209 events in the 24,255 nonstoppers was 1.13 (CI = 0.82–1.56). The number of events was insufficient to conduct meaningful subgroup analyses.



APPENDIX FIGURE 2. Distribution of eligible women by number of NHS “trials” of hormone therapy discontinuation in which they participated.

APPENDIX TABLE 6. Number of Participants, Therapy Stoppers, and CHD Events in Each NHS “Trial” to Estimate the Intention-to-Treat Effect of Discontinuation of Estrogen/Progestin Therapy

Trial	Questionnaire Yr	Participants	Stoppers	CHD Events
1	1984	107	59	2
2	1986	438	172	14
3	1988	1311	327	26
4	1990	2917	619	42
5	1992	4303	670	47
6	1994	5736	867	54
7	1996	7446	917	48
8	1998	6614	986	28

The Sound and the Fury

Was It All Worth It?

Robert N. Hoover

Abstract: The initial report of coronary heart disease (CHD) results from the trial of menopausal hormone therapy within the Women's Health Initiative precipitated substantial surprise and concern in the epidemiology research community over the apparent differences between the trial results and those of observational studies. What followed was 6 years of discussion and debate, frequently acrimonious, along with intense methodologic and substantive research attempting to reconcile or explain the apparent differences. The results have been an impressive improvement in methods to contrast and combine studies of differing designs, dramatic illustrations of some central epidemiologic principles, insights into likely mechanisms of CHD, and increasing clarity of the public health message about menopausal hormone therapy.

(*Epidemiology* 2008;19: 780–782)

The early termination of the Women's Health Initiative (WHI) trial of combined estrogen/progestin menopausal hormone therapy precipitated a profound reaction by the media, among menopausal women, and in the clinical community.¹ The trial result, that cumulative risks outweighed benefits, overturned conventional wisdom and decades of clinical practice. Particularly disturbing was the increased risk of coronary heart disease (CHD) in the face of widely accepted results from epidemiologic and clinical investigations suggesting protective effects of hormone therapy. Profound angst ensued for many epidemiologists and biostatisticians, leading to years of attempts to reconcile or explain the apparent differences between the trial results and those of the observational studies.^{2–5} The intensity of this reaction was surprising, because epidemiologists spend years learning, and careers teaching, the litany of relative strengths and weaknesses of different study designs, and the distinct possibility of differing results because of differing designs. That said, the methodologic advances and enhanced understanding of the

effects of hormone therapy resulting from this level of concern have been major contributions to our discipline.

From the methodologic perspective, investigators from the WHI⁴ and Nurses' Health Study (NHS)⁵ have proposed innovative methods for analyzing data from clinical trials and observational cohort studies in a comparable manner. This not only allows assessment of similarities and differences in findings, but more importantly, these approaches provide ways to combine data coherently from different designs. Such pooled analyses not only increase power but also can leverage the distinct strengths of each design. The result is a whole that is truly greater than the sum of its parts.

The paper by Hernán et al⁵ in this issue of the journal confirms previous analyses and speculations that many of the apparent differences in overall CHD risk between the 2 studies lie not in classic confounding by other causal variables, but in differing risks by duration of follow-up, and in groups of women defined by duration of interval between menopause and initiation of hormone therapy, along with differing distributions in these variables between the 2 study populations. Because the strongest evidence for an adverse effect occurs in the years immediately following the initiation of hormone therapy, correcting for the misclassification bias in this time interval within previous NHS analyses (associated with counting exposure from the date of return of the first questionnaire after initiation of MHT rather than an estimated time of initiation itself) further narrows apparent differences in this interval. As might be expected when subgroups of 2 studies are compared, confidence intervals are wide. Thus, comparability is hard to assess quantitatively; qualitatively, both the WHI and the NHS appear to show excess risk of heart disease immediately following initiation of hormones. Furthermore, this risk is, more pronounced among those who initiate such therapy more than a decade after menopause than in those who initiate it earlier. The cumulative excess risk dissipates with time within both studies, perhaps progressing to a decreased risk with long-term use.

Hernán et al believe that the aforementioned differences largely explain any discrepancies between the WHI and NHS results, and that residual confounding for the initiation of therapy in the NHS plays small role. The major suspect when observational study results and trial results disagree is unknown risk factors that could be controlled only by

From the Epidemiology and Biostatistics Program, Division of Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland.

Correspondence: Robert Hoover, MD, Epidemiology and Biostatistics Program, Division of Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd., EPS/8094, Bethesda, MD 20892. E-mail: hoover@mail.nih.gov.

Copyright © 2008 by Lippincott Williams & Wilkins

ISSN: 1044-3983/08/1906-0780

DOI: 10.1097/EDE.0b013e318188e21d

randomization. With regard to hormone therapy, this concern has focused primarily on potential compliance and survivor biases for longer-term users (a healthy persistent-medication-user effect).^{6,7} As Hernán et al point out, it is harder to assess the consistency of results with an assumption of no unmeasured confounders for treatment discontinuation than for initiation. If some of the remaining discrepancies (magnitude of the initial excess risk, slope of the decline in excess risk with duration of use, and the magnitude of protection, if any, with long-term use) are not simply because of chance, the possibility of unrecognized confounding remains. The plausibility of this could be contested but will probably not be resolved by further methodologic work, now that the trial is over.

Other important methodologic observations have emerged from the comparisons and contrasts associated with these reconciliation processes. Few have been new methodologic insights per se, but several are compelling examples and illustrations of established epidemiologic principles, namely the relative strengths and weaknesses of various study designs alluded to in the first paragraph. These illustrations of central principles have the potential to become particularly effective didactic tools in educating epidemiology students and colleagues from other disciplines about the richness of our discipline's strategies to contribute to understanding health and disease. For example, one weakness of a trial typically lies in the area of its generalizability. In MacMahon's words of almost 50 years ago, "the interpretation of results in terms of general applicability may be limited."⁸ Examples of this abound for therapeutic effects, which can vary by characteristics of populations chosen for study (eg, age, sex, race, and general health status). Now we have an example of an etiologic factor the effect of which apparently varies substantially with only a 10-year difference in time of exposure. Not only does this point out the potential limitations of generalizability of a trial, and the need to define from its outset the population to which you wish it to relate, but also the implications for statistical power. Trialists understandably are excited to explore subgroup effects but rarely power their studies with this in mind. This can work for a therapy trial, because another trial to explore an interesting subgroup observation can often be launched. For prevention trials, which are larger and more complex to start with, and entail added ethical issues, further trials are rarely feasible.

Interestingly, the current circumstance points out that this "weakness" may be turned to an advantage. This can happen when important biologic insights result from assessing the effect of an exposure in an understudied group. Indeed, the nonrepresentative nature of the WHI study group was recognized at the start of the trial, and a review of its protocol by a National Academy of Sciences committee suggested that the inclusion of women a decade after menopause was one of the strengths of the study design for precisely this reason.⁹ This turned out to be the case. The

differences in risk of clinical CHD by timing of exposure in the trial, and the subsequent exploration of this in the limited observational data available, has contributed directly to speculation about a mechanism that would involve increased clotting risk in women with preexisting subclinical CHD and different effects and mechanisms in those without.

Another key design effect is illustrated by the difficulty in unbiased assessment of the risks of hormone therapy shortly after beginning use (exposures beginning between follow-up efforts) in the NHS. Although a general strength of the cohort design is in unbiased exposure assessment, it frequently presents challenges in assessment of the details of timing of exposure and risk. The multiple endpoints under study, multiple risk factors being assessed, time intervals between exposure assessments, and deaths and losses to follow-up within these intervals (the probability of which may relate to exposure), all conspire to obscure detailed temporal relations between exposure and disease. This contrasts with the relative ease of collecting data in a case-control design, with its focus on 1 disease and a limited set of potential risk factors, with total exposure histories for these factors up to diagnosis. If these potential misclassification biases are an issue in even the NHS, with its compliant population and 2-year follow-up intervals, the hormone therapy story must be a cautionary tale for epidemiologists working with cohort studies that have higher drop-out rates and longer intervals between exposure updates.

When differing results emanate from different disciplines—or from different high-quality studies within the same discipline—good public health practice can be hard to define. Fortunately in this instance, the public health message for menopausal women is quite clear, despite any residual methodologic questions or differences of opinion. In addition to the patterns of risk of CHD, hormone therapy increases the risk of breast cancer, with a pattern of risk a mirror image of that for CHD—high levels of risk emerging with longer-term use, and in those initiating therapy near the time of menopause.¹⁰ Hormone therapy is also associated with increased risks of stroke,¹¹ blood clots,¹² dementia,¹³ gall bladder disease,¹⁴ and urinary incontinence¹⁵ and shows no clinically significant benefit for health-related quality of life.¹⁶ These cumulative risks so overwhelm the protective effects for osteoporosis that a clear consensus of epidemiologists and trialists has emerged. Hormonal therapy may be appropriate for the short-term treatment of moderate-to-severe vasomotor symptoms in recently menopausal women, but it should not be used long-term for the prevention of chronic disease.

In the immediate aftermath of the controversy 6 years ago, the editors of *Epidemiology* called on our community to address the issues raised by the WHI results, arguing that such an effort "cannot help but sharpen the peculiar and rigorous qualities of thinking demanded by observational

data.”¹⁷ Many of our best epidemiologists and biostatisticians have responded, and the intensity with which they have contemplated the navel of our discipline since that time has clearly achieved this goal.

ABOUT THE AUTHOR

ROBERT HOOVER is Director of the intramural Epidemiology and Biostatistics Program at the National Cancer Institute. He published the first epidemiologic study suggesting menopausal hormone therapy as a cause of breast cancer. He is continually humbled by how much he still has to learn about the epidemiologic method after nearly 40 years of trying to practice it.

REFERENCES

- Kolata G. Hormone studies: what went wrong? *The New York Times*. April 22, 2003.
- Grodstein F, Manson JE, Stampfer MJ. Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. *J Women's Health*. 2006;15:1, 35–44.
- Rossouw JE. Postmenopausal hormone therapy for disease prevention: have we learned any lessons from the past? *Nature*. 2008;83:14–16.
- Prentice RL, Langer R, Stefanick ML, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the women's health initiative clinical trial. *Am J Epidemiol*. 2005;162:404–414.
- Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19:766–779.
- Sturgeon SR, Schairer C, Brinton LA, et al. Evidence of a healthy estrogen user survivor effect. *Epidemiology*. 1995;6:227–231.
- Egeland GM, Kuller LH, Mathews KA, et al. Premenopausal determinants of menopausal estrogen use. *Prev Med*. 1991;20:343–349.
- MacMahon B. Experimental epidemiology. In: MacMahon, Pugh, Ipsen, eds. *Epidemiologic Methods*. Boston, MA: Little Brown and Co; 1960.
- Thaul S, Hotra D, eds. *An Assessment of the NIH Women's Health Initiative*. Washington, DC: National Academy Press; 1993.
- Prentice RL, Chlebowski RT, Stefanick ML, et al. Estrogen plus progestin therapy and breast cancer in recently postmenopausal women. *Am J Epidemiol*. 2008;167:1207–1216.
- Wassertheil-Smoller S, Hendrix SL, Limacher M, et al. Effect of estrogen plus progestin on stroke in postmenopausal women. *JAMA*. 2003;289:2673–2684.
- Cushman M, Kuller LH, Prentice R, et al. Estrogen plus progestin and risk of venous thrombosis. *JAMA*. 2004;292:1573–1580.
- Shumaker SA, Legault C, Rapp SR, et al. Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women. *JAMA*. 2003;289:2651–2662.
- Cirillo DJ, Wallace RB, Rodabough RJ, et al. Effect of estrogen therapy on gallbladder disease. *JAMA*. 2005;293:330–339.
- Hendrix SL, Cochrane BB, Nygaard IE, et al. Effects of estrogen with and without progestin on urinary incontinence. *JAMA*. 2005;293:935–948.
- Hays J, Ockene JK, Brunner RL, et al. Effects of estrogen plus progestin on health-related quality of life. *N Engl J Med*. 2003;348:1839–1854.
- Editors. Epidemiology and randomized clinical trials. *Epidemiology*. 2003;14:2.

ITT for Observational Data

Worst of Both Worlds?

Meir J. Stampfer

Abstract: Hernán et al reanalyzed Nurses' Health Study Data on hormone therapy and heart disease, to explore further the apparent discrepancy for those results compared with findings from the Women's Health Initiative Trial. Hernán et al concludes that differences in time since menopause remains the most plausible explanation for the different findings. Part of the analysis employs application of the "intention-to treat" principle to analyze the observational data. This commentary points out some of the weaknesses inherent in that approach, which combines a major limitation of observational studies—lack of randomization—with a common limitation of trials, imperfect adherence to the assigned treatment.

(*Epidemiology* 2008;19: 783–784)

In this issue, Hernán et al provide a reanalysis of Nurses' Health Study (NHS) data to explore reasons for the apparent discrepancy in results for use of estrogen plus progestin postmenopausal hormones and heart disease, which were associated with lower risk, in contrast to the positive association in the Women's Health Initiative (WHI) randomized trial. Their main conclusion was that most of the difference could be attributed to the difference in age distribution at the time of initiation of hormone therapy in these 2 studies. Consistent with prevailing clinical practice, most hormone users in NHS began around menopause, whereas in the WHI, two thirds of the participants were aged 60 years or older at the start. The hypothesis that the time of initiation of hormone therapy affects risk of coronary disease—the timing hypothesis—was explored by Grodstein et al¹ soon after publication of the initial WHI findings. Since then, considerable additional evidence, based on animal and human studies, has accumulated to support this hypothesis, as reviewed by Manson and Bassuk² and Mendelsohn and Karas.³

The NHS and WHI results have been remarkably concordant for all other clinical outcomes examined, including

stroke, pulmonary embolism, breast cancer, and colorectal cancer. In particular, the virtually identical results for stroke, as confirmed by the most recent NHS analysis⁴ argue against substantial confounding by lifestyle factors or other variables in NHS analyses. Hernán et al came to a similar conclusion, which is that residual confounding is unlikely to explain the apparent divergent findings.

Consistent evidence suggests a transient increase in risk among women who start hormone therapy years after menopause, in the presence of preexisting atherosclerosis, but not in women who start earlier. Further evidence suggests subsequent protection, reflected by the finding that in the WHI the cumulative incidence curves converged by 8 years. Using the retrospective data in the NHS, Hernán et al show this transient increase in risk among women who started hormone therapy more than 10 years after menopause but not in those who started earlier. This was also examined in a sensitivity analysis in our earlier publication; in agreement with Hernán et al, this transient increase in risk did not have a substantial impact on the overall findings in NHS.

In a novel approach to considering this issue, Hernán et al apply the intention-to-treat (ITT) principle in analyzing the NHS data, to "mimic the design of the randomized trial as closely as possible in the NHS." The main advantage of a randomized trial is, of course, that potential confounding factors, both known and unknown, will tend to be evenly distributed across groups. This advantage is lost if only adherent individuals are considered in the analysis, so the ITT principle is appropriately typically applied to analyze all data regardless of adherence to randomized treatment assignment. However, as a consequence, poor adherence in a randomized trial will tend to yield inaccurate estimates of the efficacy of the treatment under study due to misclassification of exposure. In observational data, there is concern about the potential for confounding, but with repeated assessments, the actual exposure of the individuals can be identified, reducing misclassification. In applying the ITT principle to observational data, those who initiate the exposure (in the present example, hormone use) are treated as though they were in that exposure group, regardless of their actual later behavior—that is, if they stopped such use. Thus, application of the intention-to-treat principle to observational data essentially combines the most important limitations of each study design.

From the ^aDepartment of Epidemiology, Harvard School of Public Health and ^bChanning Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.

Supported by NIH grants HL080644 and CA87969.

Correspondence: Meir J. Stampfer, Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115. E-mail: mstampfer@hsph.harvard.edu.

Copyright © 2008 by Lippincott Williams & Wilkins

ISSN: 1044-3983/08/1906-0783

DOI: 10.1097/EDE.0b013e318188442e

With the resulting misclassification of exposure, it is no surprise that essentially null results emerge from the present ITT analysis. This does not “explain” the apparent discrepancy, it just tells us that substantially misclassified exposure data will tend to yield null results. It is a kind of magical thinking that by using the terminology of randomized trials and calling the 2-year intervals of observation “trials,” the advantages of randomization are achieved. To the contrary, the worst features of both designs emerge.

Hernán et al attempt to deal with this misclassification by adjusting the results for adherence by using inverse probability weighting. This makes an analysis that is already complex and difficult to follow even more complicated by requiring additional assumptions and models. The magnitude of this problem is apparent by the finding that the relative risk for 8+ years was 0.87 in Hernán’s unadjusted analysis and only changed to 0.85 in the adherence-adjusted analysis, despite the fact that over half of the women were misclassified by that time in the ITT analysis.

One potential use of an ITT analysis might be to evaluate whether the women who discontinue their hormone use are at elevated risk, which could lead to an apparent lower risk for those who continue. However, this can also be evaluated using conventional methods by examining risk in

those who discontinue their exposure. This is an interesting attempt to use a novel method for analysis, but one that adds no new insights on the relation of hormone therapy to chronic heart disease. Because of its far greater complexity and “black box” nature, which make it difficult even to track numbers of subjects, it should not be recommended for routine use.

ABOUT THE AUTHOR

MEIR STAMPFER is Professor of Epidemiology and Nutrition at Harvard School of Public Health, where he served as chair of the Department of Epidemiology, from 2000 to 2007. He is also Professor of Medicine, and serves as Associate Director of Channing Laboratory at Brigham and Women’s Hospital. In addition to his research in chronic disease epidemiology, he also directs 2 training grants.

REFERENCES

1. Grodstein F, Clarkson TB, Manson JE. Understanding the divergent data on postmenopausal hormone therapy. *N Engl J Med*. 2003;348:645–650.
2. Manson JE, Bassuk SS. Invited commentary: hormone therapy and risk of coronary heart disease—why renew the focus on the early years of menopause? *Am J Epidemiol*. 2007;166:511–517.
3. Mendelsohn ME, Karas RH. HRT and the young at heart. *N Engl J Med*. 2007;356:2639–2641.
4. Grodstein F, Manson JE, Stampfer MJ, Rexrode K. Postmenopausal hormone therapy and stroke: role of time since menopause and age at initiation of hormone therapy. *Arch Intern Med*. 2008;168:861–866.

Data Analysis Methods and the Reliability of Analytic Epidemiologic Research

Ross L. Prentice

Abstract: Publications that compare randomized controlled trial and cohort study results on the effects of postmenopausal estrogen-plus-progestin therapy are reviewed. The 2 types of studies agree in identifying an early elevation in coronary heart disease risk, and a later developing elevation in breast cancer risk. Effects among women who begin hormone therapy within a few years after the menopause may be comparatively more favorable for coronary heart disease and less favorable for breast cancer. These analyses illustrate the potential of modern data analysis methods to enhance the reliability and interpretation of epidemiologic data.

(*Epidemiology* 2008;19: 785–788)

There is a pressing need to assess and enhance the reliability of findings from observational studies. Available methods for controlling confounding, measurement error and other biases can provide adjustments in the desired direction, but objective means of assessing bias avoidance are generally lacking. Randomized controlled trials include an objective assignment of a study treatment or intervention and avoid confounding by prerandomization factors. However, trials are expensive, and typically cannot be conducted in a manner that powerfully addresses subset hypotheses or treatment effects over long periods of exposure. Hence, the population science research agenda must rely heavily on observational studies for the development and initial testing of disease prevention hypotheses, with trials typically conducted only for well established hypotheses that have strong public health potential.

Settings in which both trials and observations studies are available provide a particular opportunity to examine consistency of results from the 2 types of studies, and to identify improvements in study design, conduct, or analysis that may help to explain any discrepancy in results. Such data

exist for postmenopausal hormone therapy in relation to several important clinical outcomes. Few topics have generated more interest and controversy in recent years, in part because findings from a clinical trial and observation studies seemed to be strongly discrepant.

Benefits and Risks of Postmenopausal Hormone Therapy

A substantial body of cohort and case-control studies has suggested that postmenopausal hormone therapy reduces coronary heart disease (CHD) risk by about 40% to 50%, with little indication for a difference in effects between estrogen-alone or estrogen-plus-progestin.^{1,2} A subsequent and extensive observational literature has also suggested elevations in breast cancer risk, by about 30% for estrogen and 50% to 100% for estrogen-plus-progestin.^{3,4} Reports available by the early 1990s informed the design of the Women's Health Initiative (WHI) clinical trial, which randomized 10,739 posthysterectomy women to 0.625 mg daily conjugated equine estrogen, and 16,608 women with intact uterus to this same estrogen regimen plus 2.5 mg/d medroxyprogesterone acetate. CHD was the designated primary outcome with breast cancer as the primary "safety" outcome in both trials. A recruitment age range of 50 to 79 years was specified to examine whether health benefits and risk would apply broadly to postmenopausal women. At the time these trials were initiated the estrogen and estrogen-plus-progesterone regimens under study were used by about 8 and 6 million women, respectively, in the United States.

With this background it came as quite a surprise when the estrogen-plus-progesterone trial was stopped prematurely in 2002. Health risks were judged to exceed benefits over its 5.6-year average follow-up period. The health risks included elevations in breast cancer, stroke, venous thromboembolism, and CHD, which were only partially offset by reductions in fractures and colorectal cancer.⁵ Although breast cancer was a trigger for early stopping, the hazard ratio (HR) estimate was a moderate 1.24 with 95% confidence interval (CI) = 1.01–1.54.⁶ More surprising was the HR of 1.24 (95% CI from 1.00–1.54) for CHD, with an HR of 1.81 (95% CI = 1.09–3.01) during the first year of combined-hormone use.⁷

WHI investigators undertook joint analyses of data from the clinical trial with data from a corresponding subset

From the Fred Hutchinson Cancer Research Center, Seattle, WA.
Supported by NCI grant P01 CA53996. The WHI Clinical Coordinating Center is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, US Department of Health and Human Services through Contract N01WH22110.

Correspondence: Ross L. Prentice, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M3-A410, POB 19024, Seattle, WA 98109-1024. E-mail: rprentic@fhcrc.org.

Copyright © 2008 by Lippincott Williams & Wilkins

ISSN: 1044-3983/08/1906-0785

DOI: 10.1097/EDE.0b013e318188e83b

of the WHI observational study, which was composed of women recruited from the same population as the trial, with much commonality in eligibility criteria, baseline data collection, and outcome ascertainment. HRs from the observational study alone were considerably lower than for the trial and similar to those from other cohort studies following confounding control, for each of CHD, stroke, and thromboembolism.⁸ However, HRs for CHD agreed closely following control for time from hormone therapy initiation (duration of use among adherent women). The same analytic techniques did not seem to explain fully the lower risk of stroke in the observational study compared with the trial.

The WHI estrogen trial was also ended early (in 2004), based on an elevation in stroke risk similar to that for estrogen-plus-progesterone (HR ~1.3), and a limited power to establish a CHD effect before the trial's planned termination.⁹ The HR (95% CI) for CHD was 0.95 (0.79–1.15), whereas that for breast cancer was a rather surprising 0.80 (0.62–1.04) over the trial's follow-up period that averaged 7.1 year.^{10,11}

Comparative analyses of WHI trial and observational study data for CHD, stroke, and thromboembolism yielded almost identical results for estrogen as for estrogen-plus-progesterone. HRs from the 2 sources agreed closely for CHD and thromboembolism, and not so closely for stroke, after confounding control on allowing the HR to depend on time from estrogen initiation.¹² In fact, the ratio of HRs from the trial and the observational study was about 0.9 for CHD and thromboembolism, and about 0.7 for stroke for both hormone regimens, presumably suggesting some residual bias for stroke.

Comparative analyses of this type were recently presented for breast cancer.^{13,14} HRs from the observational study were somewhat higher than those from the trial for both hormone regimens, even after control confounding and accommodating time since initiation of hormone therapy. These HRs, however, were higher among women who first initiated hormones within a few years after menopause compared with women having larger gap times. The HRs agreed closely between the 2 data sources after allowing effect modification by this gap time variable. Among women having gap times of less than 5 years the breast cancer HR increased to about 2.0 following 2 or more years of estrogen-plus-progesterone, whereas that for estrogen alone was about 1.0.

The types of modeling and comparative analyses just described achieve some robustness by virtue of similar findings between the 2 hormone regimens, but it is also of great interest to compare the clinical trial results with results from other observational studies, including the Nurses Health Study (NHS), which played an important role in the generation and initial testing of hypotheses related to hormone treatment effects.

Estrogen-Plus-Progesterone Therapy and CHD in the NHS

In this issue Hernán et al¹⁵ provide a reanalysis of the association between estrogen-plus-progesterone and CHD in the NHS. These authors are to be congratulated on a careful matching of the NHS subset used (34,575 women) to the set of women enrolled in the WHI trial of combined hormones, and for a series of analyses that elucidate the impact of various analytic definitions and estimation procedures on the resulting HRs. Also, the participating NHS coauthors are to be congratulated for allowing their data to be subjected to these novel analytic approaches. Compared with the WHI observational study, the NHS has the distinct advantage that much of the hormone use was initiated after cohort enrollment, potentially allowing precise assessment of benefits and risks during the early months after hormone initiation. Previous analyses of NHS data relied on a biennial snapshot of current hormone user status. This was evidently an important analytic limitation for estimation of an early HR increase that substantially dissipated within a year or 2 after initiation of hormone therapy. For example, women who initiate hormones would be classified based on this snapshot as nonusers until their biennial data collection time, and permanently as nonusers if they stopped usage before such collection.¹⁶ In the present analysis, the authors recover "estimates" of the date of hormone initiation to the extent possible through a fuller use of available data, presumably substantially mitigating this source of bias. They also attempt to emulate a clinical trial by defining a multivariate response for each woman. Each woman is classified as an initiator or noninitiator in each 2-year follow-up interval. An initiator versus noninitiator HR is then estimated from the follow-up of each such "stratum" with appropriate provision for dependencies that arise from an individual woman contributing to several (up to 8) HR estimates. There was little evidence that such HR estimates differed among strata, and the resulting common HR estimates agreed closely with corresponding estimates from the WHI trial with HR estimate (95% CI) of 1.42 (0.92–2.20) for the first 2 years of use, and 0.96 (0.78–1.18) over the entire follow-up period. Note that this type of methodology for emulating clinical trials was not needed for analysis of the WHI observational study because there were few hormone initiators after cohort enrollment.

Hernán et al go on to describe a possible interaction ($P = 0.08$) of HR with years from menopause to initiation of hormone therapy. Among women having fewer than 10 years from menopause to initiation of hormones, the HR (95% CI) was 1.28 (0.62–0.84) in the first 2 years of follow-up, and 0.81 (0.56–1.17) thereafter. Such an interaction was not evident in the WHI trial and would benefit from study in other settings.

Intention-to-Treat and Adherence Adjustment

Hernán et al include some rather harsh criticisms of intention-to-treat (ITT) analyses, indicating that ITT estimates “may be unsatisfactory when studying efficacy, and inappropriate when studying the safety, of an active treatment compared with no treatment.” It seems worth reiterating that of the various analyses discussed here, it is only for the ITT comparisons in the clinical trial that we can be sure the treated and untreated groups were fully comparable at enrollment. Hence, if the clinical outcomes are equally ascertained between the active and placebo groups, a causal interpretation for the treatment and its sequelae is justified for any differences that emerge. By comparison, what Hernán et al refer to as an ITT analysis of the NHS data attempts to argue toward a causal interpretation by virtue of careful confounding control, and accommodation of time of hormone initiation, and time since hormone initiation. There is limited ability in the absence of corresponding clinical trial data to assess the success of these efforts.

However, there are important questions to answer beyond ITT comparisons. One is the magnitude of treatment effects among study subjects who adhere to the treatment regimen. Even the clinical trial setting does not allow an estimate of risk for adherent women without making additional assumptions. Women who adhere to treatment or nontreatment status may have many biobehavioral differences from those who do not, and these characteristics may differ between treated and nontreated groups. A trial that is able to maintain an effective blinding of active versus placebo status may yield fairly comparable groups of adherent women.⁵ Nevertheless, WHI investigators describe comparisons between women adherent to active and placebo pill-taking as “sensitivity analysis,” to alert the reader to possible noncomparability between these groups.

Some adherence-adjusted analyses in WHI have simply censored the follow-up of women soon after they become nonadherent. Including inverse censoring probability weighting as in Hernán et al, could presumably enhance these comparisons by restoring a contrast that is theoretically applicable to the entire randomized group. Although this method is a useful step forward, the justification for the adherence-adjusted HRs that emerge depends directly on the ability to model the nonadherence process. Doing so is analogous to modeling for control of confounding. The factors that determine adherence to each treatment group in the study population must be accurately measured and correctly modeled. It would seem that the knowledge base for this type of analysis is still limited, arguing for a suitably circumspect interpretation of resulting HR estimates. HRs among adherent women tend to be more extreme in their departure from the null than do ITT analyses for both the cardiovascular disease and breast cancer outcomes for each of the data sources considered here.

Excellent progress has been made in recent decades on the development of data analytic methods for trials and observational studies emanating, in part, from the Cox¹⁷ HR regression model and its multivariate extensions. The reanalysis by Hernán et al strongly suggests that the use of these methods can strengthen the analysis and interpretation of observational studies. Still, it seems evident that clinical trials are needed when preventive interventions are widely used or when the public health implications are sufficiently large. In the special case of postmenopausal hormone therapy, the state of knowledge of health benefits and risks is quite different following the WHI trials than had been assumed in advance. It is interesting to question whether an early elevation in CHD risk, or a more sustained elevation in stroke and dementia risk,^{18–20} would have been identified in the absence of clinical trial data.

REFERENCES

- Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med.* 1991;20:47–63.
- Grady D, Rubin SM, Petitti DB, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med.* 1992;117:1016–1036.
- Beral V, Million Women Study Collaborators. Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet.* 2003; 362:419–427.
- Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet.* 1997;350: 1047–1059.
- Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA.* 2002;288:321–333.
- Chlebowski RT, Hendrix SL, Langer RD, et al. Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women: the Women's Health Initiative randomized trial. *JAMA.* 2003;289:3243–3253.
- Manson JE, Hsia J, Johnson KC, et al. Estrogen plus progestin and the risk of coronary heart disease. *N Engl J Med.* 2003;349:523–534.
- Prentice RL, Langer R, Stefanick ML, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *Am J Epidemiol.* 2005;162:404–414.
- Anderson GL, Limacher M, Assaf AR, et al. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. *JAMA.* 2004; 291:1701–1712.
- Hsia J, Langer RD, Manson JE, et al. Conjugated equine estrogens and coronary heart disease: the Women's Health Initiative. *Arch Intern Med.* 2006;166:357–365.
- Stefanick ML, Anderson GL, Margolis KL, et al. Effects of conjugated equine estrogens on breast cancer and mammography screening in postmenopausal women with hysterectomy. *JAMA.* 2006;295:1647–1657.
- Prentice RL, Langer RD, Stefanick ML, et al. Combined analysis of Women's Health Initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *Am J Epidemiol.* 2006;163:589–599.
- Prentice RL, Chlebowski RT, Stefanick ML, et al. Estrogen plus progestin therapy and breast cancer in recently postmenopausal women. *Am J Epidemiol.* 2008;167:1207–1216.
- Prentice RL, Chlebowski RT, Stefanick ML, et al. Conjugated equine

- estrogens and breast cancer risk in the Women's Health Initiative clinical trial and observational study. *Am J Epidemiol*. 2008;167:1407–1415.
15. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19:766–779.
 16. Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative (with discussion). *Biometrics*. 2005; 61:899–941.
 17. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc B*. 1972;34:187–220.
 18. Wassertheil-Smoller S, Hendrix SL, Limacher M, et al. Effect of estrogen plus progestin on stroke in postmenopausal women: the Women's Health Initiative: a randomized trial. *JAMA*. 2003;289:2673–2684.
 19. Hendrix SL, Wassertheil-Smoller S, Johnson KC, et al. Effects of conjugated equine estrogen on stroke in the Women's Health Initiative. *Circulation*. 2006;113:2425–2434.
 20. Shumaker SA, Legault C, Rapp SR, et al. Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *JAMA*. 2003;289:2651–2662.

A Call for Nominations: The 2009 Rothman EPIDEMIOLOGY Prize

EPIDEMIOLOGY presents an annual award for the best paper published by the journal during the previous year. The prize of \$3,000 and a plaque goes to the author whose paper is selected by the Editors and the Editorial Board for its originality, importance, clarity of thought, and excellence in writing.

With this issue, we close our 2008 volume. We invite our readers to nominate papers published during the past year. Please email your nominations to Allen Wilcox, Editor-in-Chief: EDITOR@EPIJOURNAL.ORG.

Nominations must be received no later than 31 December 2008. The winner will be announced in our September 2009 issue and at the annual meeting of the American College of Epidemiology.

This award is made possible by an endowment from Hoffman-LaRoche Ltd., managed by the American College of Epidemiology.

Authors' Response, Part I: Observational Studies Analyzed Like Randomized Experiments

Best of Both Worlds

Miguel A. Hernán^a and James M. Robins^b

We thank the 3 discussants for their contributions. The Women's Health Initiative (WHI) investigators found a greater CHD risk in the estrogen plus progestin (ie, combined hormone) therapy arm than in the placebo arm of the trial (hazard ratio [HR]: 1.24, 95% confidence interval [CI]: 1.00–1.54).¹ In contrast, the Nurses' Health Study (NHS) investigators found a lower CHD risk in current users of combined hormone therapy than in never users (HR: 0.68, 95% CI: 0.55–0.83, in their most recent publication).² We investigated possible reasons for this discrepancy by reanalyzing the NHS data; we used a novel approach that conceptualizes a follow-up observational study as a sequence of "trials."³ The discussants disagree sharply in their assessments of the value of this analytic strategy. Prentice⁴ and Hoover⁵ are positive, whereas Stampfer⁶ finds that our approach combines the limitations of both observational studies and randomized trials, gives biased adherence-adjusted HR estimates, and "adds no new insights on the relation of hormone therapy to CHD." He criticizes our approach for its complexity, its need for additional assumptions, and its "black box" nature, and argues for the continued use of the conventional methods routinely employed in NHS publications, owing to their transparency and validity. We now address each of these criticisms.


New Insights

Consider the question of "new insights." Our reanalysis suggests there is a short-term increase in CHD incidence after initiation of combined hormone therapy among all NHS women. Furthermore, our reanalysis suggests effect modification of the HR for combined hormone therapy by years since menopause (*P* value: intention-to-treat 0.08, adherence-adjusted 0.01). Neither of these results is found in any previous NHS analyses. On the other hand, these results are consistent with the WHI estimates¹ although, as pointed out by Prentice⁴ the *P* value for interaction is <0.05 in the WHI only when time since menopause is coded as an ordered categorical variable.⁷

Thus our analyses contain 3 new insights. First, discrepancies between previous NHS and WHI results in regard to the 2 results above appear to be due to the NHS analytic approach and not to any inherent problems in the NHS data. Second, these discrepancies disappear when our approach is used to analyze the NHS data. Third, our results are consistent with the so-called "timing hypothesis," which states that the increased CHD risk is concentrated in women who start combined hormone therapy many years after menopause. It is perplexing that, in his discussion of the timing hypothesis, Stampfer does not mention either the evidence provided by our paper, or the relative lack of evidence in

From the ^aDepartment of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts; ^bDepartments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts.

Supported by National Institutes of Health grant HL080644.

 Supplemental material for this article is available with the online version of the journal at www.epidem.com; click on "Article Plus."

Copyright © 2008 by Lippincott Williams & Wilkins

ISSN: 1044-3983/08/1906-0789

DOI: 10.1097/EDE.0b013e318188e85f

a recent NHS paper² he coauthored, in which the HR estimates were <1 both in women near menopause and >10 years from menopause.

Our reanalysis of the NHS reconciles these previously discrepant results, but leaves important questions unanswered: the effect modification by time-since-menopause was found in both the WHI and the NHS, but was not found by us in a British population in which a different type of combined hormone therapy was used.⁸ Also, our reanalysis, but not the WHI, suggested effect modification by age.³ Finally, the apparent discrepancies between WHI estimates and previous NHS estimates were most noticeable for CHD and overall mortality; our reanalysis has focused only on CHD.

Bias Rebuttal

Consider next Stampfer's claim that there is empirical evidence of bias in our adherence-adjusted HR estimates. He argues that the adherence-adjusted HR estimate of 0.85 in the 8+ years stratum must be wrong, because it differs only negligibly from the unadjusted (ie, intention to treat [ITT]) estimate of 0.87, and yet most stratum members were non-adherent. We make 2 points in rebuttal: the first a bit subtle and the other not. The adherence-adjusted effect will be less than the unadjusted effect only under the hypothesis that combined hormone therapy is protective. Thus, Stampfer's claim requires that one already accept as true the very hypothesis being examined. Second, Stampfer considers only the point estimate of the adherence-adjusted effect, and not the very wide associated 95% CI (0.22–3.19). The lower confidence limit of 0.22 is clearly less than the unadjusted estimate of 0.87. In fact, it is because of the large degree of uncertainty in the 8+ year stratum that we restricted the survival curves in Figures 1 and 3 to the first 8 years of follow-up.

Whose Black Box

To align our discussion with Stampfer's concerns, we henceforth take as our inferential goal the estimation of the effect of continuous combined hormone therapy on CHD risk in postmenopausal NHS women. The conventional NHS analytic method compares, at each time, the 2-year risk of CHD among currently exposed women with that of women never exposed, adjusting for the current (updated) values of the potential confounders. In our view, it is the conventional method, not ours, that is the black box and potentially biased. To justify our claim, we first describe the considerations that led us to our chosen analytic strategy. We then compare our adherence-adjusted HR estimate of the effect of continuous hormone therapy with conventional NHS estimates.

Possible Explanations

A number of different biologic and methodologic explanations for the difference between the WHI and NHS estimates have been proposed. The leading biologic explanation is the timing hypothesis. The most common method-

ologic explanations⁹ are that the NHS observational estimates are biased for one or more of the following reasons:

- Healthy initiator bias: women initiating combined hormone therapy are at lower risk of CHD than noninitiators (and thus not comparable with the noninitiators), even within levels of the covariates being adjusted for.
- Healthy continuer bias: among women using combined hormone therapy, those who continue therapy are at lower risk of CHD than those who stop therapy (and thus not comparable with those who stop), even within levels of the covariates being adjusted for.
- Misclassification of the hormone exposure of initiators. The NHS analyses updated hormone status at the time of questionnaire return, resulting in misclassification of exposure during the period from initiation of therapy to questionnaire return, a period of up to 2 years.

To assess the importance of these 3 explanations, we used the NHS data and the published data on ITT from the WHI. First, we eliminated, to the extent possible, the misclassification bias (c) by beginning follow-up at the estimated date of therapy initiation, as described in the paper.

Consider next the bias due to explanations (a) and (b). Because the noncomparability in those explanations represents confounding by unmeasured factors, one might guess that the bias in the adherence-adjusted HR attributable to this noncomparability would not be empirically estimable. Surprisingly, as described next, the bias can be empirically estimated with (and essentially only with) an ITT analysis under some often reasonable additional assumptions.

ITT Estimates of Bias

We set out to quantify the degree of healthy initiator bias (a) in the NHS by conceptualizing the NHS as a sequence of "trials" of therapy initiation. Specifically, if the healthy initiator bias is small, then within strata of those baseline risk factors with different distributions in the WHI and NHS (eg, years from menopause), the ITT effect of therapy initiation on CHD should be similar in the WHI and NHS trials, provided that the 2 studies also have similar rates of (and reasons for) nonadherence. The rate of noncompliance at 6 years was 42% (WHI) versus 55% (NHS) in initiators and 11% (WHI) versus 13% (NHS) in the noninitiators.¹⁰ The relatively close agreement between the ITT results in the NHS and WHI reported in our paper is consistent with minimal healthy initiator bias (a) and with success in eliminating most misclassification (c) in the NHS. However, the evidence just cited is not as strong as it might appear, because the definitions of nonadherence in the WHI and in our paper are not the same. We hope to apply a single definition of nonadherence to both the NHS and WHI in the future.

We next set out to quantify the healthy continuer bias (b) in the NHS by conceptualizing the NHS as a sequence of trials of therapy discontinuation (this is described in Appendix A5 of our paper³). Specifically, the ITT HR estimate of

1.13 (95% CI = 0.82–1.56) for discontinuers versus continuers can be used to estimate an upper bound on the healthy continuer bias: the difference of 0.13 by which the ITT HR exceeds 1 estimates an approximate upper bound on the relative bias in the adherence-adjusted HR attributable to healthy continuer bias, except when the true effect of therapy on CHD is both deleterious and of substantive importance.

When as in the NHS, few subjects continue on therapy for prolonged periods, the difference between the approximate bound and the actual bias will be small, even when therapy is either moderately deleterious or moderately beneficial. As a result, the worse the adherence, the better our ITT estimate of the healthy continuer bias!

Thus, under the assumption encoded in explanation (b) that women continuing therapy are healthier than those discontinuing, our ITT analysis succeeded in (approximately) bounding the healthy continuer bias without requiring a randomized trial of hormone discontinuation for comparison. By an analogous argument, an approximate upper bound on the healthy initiator bias could be derived from our ITT analysis of therapy initiation in the NHS, even without the WHI trial of hormone initiation for comparison. The critical role of these ITT analyses in producing evidence against explanations (a), (b), and (c) is not addressed by Stampfer in his discussion.

Estimates of the Effect of Continuous Treatment

As discussed above, when follow-up begins at the estimated date of therapy initiation, any bias attributable to explanations (a), (b), and (c) is likely small. We therefore used our adherence-adjusted HR estimator to estimate the effect of continuous combined hormone therapy on CHD under the assumption that all 3 explanations are false. This assumption implies that confounding by unmeasured factors and misclassification of therapy are both absent. Furthermore this assumption guarantees that our adherence-adjusted HR estimator is unbiased for the causal effect of continuous therapy, provided our models for CHD and hormone initiation/discontinuation (used in estimation of the inverse probability weights) are correctly specified.

In contrast, even in the absence of model misspecification, this assumption does not suffice to ensure that the conventional NHS analytic approach is unbiased. In fact, as discussed in our paper, conventional NHS estimates are guaranteed to be unbiased only when the measured time-dependent confounders for therapy initiation and discontinuation are not themselves affected by hormone use. Otherwise, the conventional NHS estimates may be biased in any direction, either toward or away from the null. In the electronic Appendix we provide both a heuristic explanation and a graphical proof of the bias. Thus our approach requires fewer prior assumptions for its validity than the conventional NHS approach, despite Stampfer's statement to the contrary. In fact,

even Stampfer's statement that our approach uses more modeling assumptions is inaccurate. The sole non-conventional assumption we use is the (empirically testable) assumption that our model for hormone initiation/discontinuation is nearly correct; symmetrically, as documented below, the conventional NHS analysis uses modeling assumptions that we do not.

It follows that if past use of combined hormone therapy affects some current (updated) covariate, the only strictly valid approach to determining whether a conventional NHS effect estimate is biased (and, if so, the direction and magnitude of that bias) is by comparing the NHS estimate to some known unbiased estimate of the effect of continuous exposure (ie, our adherence-adjusted HR estimate or another of the so-called g-method estimates listed in the Appendix).

We have made a large number of such comparisons. We find that, in the absence of model misspecification, conventional estimates are generally only slightly biased except when (eg, in observational studies of the effect of antiretroviral therapy on time to AIDS or death) there is strong confounding by time-dependent covariates (eg, CD4 cell count) and past treatment has a sizable effect on the covariates.

Turn now to precision. All CHD events in current users contribute to the conventional NHS estimates. In contrast, only CHD events in continuous users contribute to our adherence-adjusted estimates. Therefore, our adherence-adjusted estimates are less precise than conventional NHS estimates (compare the standard errors in Table 5 of our paper with those in the last 5 columns of the 0–24 months row in Table 6). However, with regard to the effect of continuous treatment, the greater precision of the NHS analysis comes at the price of an additional modeling assumption—the assumption that the HR among all current users is equal to the HR among the continuous users.

We therefore expected that, unless this additional modeling assumption was grossly wrong, conclusions concerning the long-term effect of continuous therapy based on a conventional NHS analysis would be consistent with, but more precise than, those based on our adherence-adjusted analysis. Although this expectation was largely fulfilled, there is an apparent exception regarding effect modification by years since menopause. Specifically, we obtained an HR of 1.20 (0.78–1.84) in women >10 years from menopause, an HR of 0.54 (0.19–1.51) in women <10 years from menopause, and an interaction *P* value of 0.15 when we fit separate models to women <10 years and to women >10 years from menopause. To increase the power to detect an interaction, we also added a single product term (the indicator for combined hormone therapy times the indicator for <10 years from menopause) to the model for the overall HR, and tested the hypothesis that the coefficient of this product term was zero. The interaction *P* value from this more powerful test was 0.01. In contrast, the conventional NHS analysis reported by Stampfer et al found HRs of 0.90 (0.62–1.29) in women >10 years and 0.71 (0.56–0.89) in women <4 years from menopause with an

interaction P value of 0.28 (calculated from the authors' table 2 in reference 2) based on separate models in women <4 and >10 years from menopause. Interaction P values from a more powerful test were neither reported nor calculable. The authors chose <4 years to define a woman "near menopause" because they "believed a 6 years cut-off was too long;" they did not indicate if this belief was based on empirical evidence in the NHS data. At present, we do not know whether this apparent exception is a consequence of a larger than expected bias in the conventional estimate of the interaction, other unexplored difference in analytic detail, or sampling variability. Regardless, the fact remains that conventional estimates require vetting by comparison with g-method estimates before they can be trusted.

Hazard Ratio versus Survival Curves

A particular example of selection bias due to conditioning on variables affected by treatment can occur when the hazard ratio is chosen as the effect measure. As an example, suppose combined hormone therapy causes a CHD event within 2 years among a substantial fraction of the women who have underlying undiagnosed coronary atherosclerosis at the time of therapy initiation, and has no effect among other women. Thus combined hormone therapy benefits no one. The average HR will be greater than 1 during the first 2 years after initiation, and less than 1 after the first 2 years. This is so because most of the hormone-exposed susceptible women will have developed CHD within 2 years of initiation and thus are removed from the calculation of the hazard ratio after year 2. In contrast, many unexposed susceptible women will survive CHD-free for 2 years; their later CHD events will contribute to the post year-2 HR. However, the survival (or, symmetrically, the cumulative incidence) curves for treated and untreated women would separate at the start of follow-up and slowly converge over time but never cross.^{11,12}

Whenever the survival curves fail to cross, it is always possible that therapy benefits no one, even if the time-specific HR is less than 1 for most of the follow-up period. Thus Stampfer is incorrect in saying that convergence of the WHI curves⁽¹⁾ suggests protection, unless there were clear statistical evidence that the curves actually crossed. Such evidence is lacking in the WHI publications.

Next consider an analysis that excludes the first 2 years of follow-up. Such an analysis implicitly conditions on surviving without CHD for 2 years, an event affected by prior therapy. It follows that one would estimate an average HR <1 and could erroneously conclude that therapy is beneficial.

This discussion is directly relevant to the interpretation of NHS data. As discussed in our paper, previous NHS analyses begin follow-up at the time of questionnaire return, which excludes from the analyzed person-time of hormone users an initial post initiation period of up to 2 years. Hence, the previous NHS analyses implicitly condition on surviving without CHD to the next questionnaire, an event affected by treatment.

In fact, the single most important difference between our adherence-adjustment approach and the conventional approach may be that we formulated the question of interest explicitly "what would be the relative risk of CHD comparing hormone therapy initiators and non initiators had they adhered to their initial treatment status during the entire follow-up?" Our methodology, including our attempts to start follow-up at the time of therapy initiation rather than at the time of questionnaire return, follows naturally from asking this question. Thus, contrary to Stampfer's assertion that our approach combines the worst limitations of both randomized trials and observational studies, in fact our analysis combines the strengths of one of the best available observational studies—large sample size, long follow-up, multiple longitudinal measurements—with a major strength of randomized trials—a well defined scientific question.

REFERENCES

1. Manson JE, Hsia J, Johnson KC, et al. Estrogen plus progestin and the risk of coronary heart disease. *N Engl J Med*. 2003;349:523–534.
2. Grodstein F, Manson JE, Stampfer MJ. Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. *J Womens Health (Larchmt)*. 2006;15:35–44.
3. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19:766–779.
4. Prentice R. Data analysis methods and the reliability of analytic epidemiologic research. *Epidemiology*. 2008;19:785–788.
5. Hoover RN. The Sound and the Fury: Was It All Worth It? *Epidemiology*. 2008;19:780–782.
6. Stampfer MJ. ITT for observational data - worst of both worlds? *Epidemiology*. 2008;19:783–784.
7. Manson JE, Bassuk SS. Invited commentary: hormone therapy and risk of coronary heart disease—why renew the focus on the early years of Menopause? *Am J Epidemiol*. 2007;166:511–517.
8. Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. *Biometrics*. 2005;61:922–930.
9. Grodstein F, Clarkson TB, Manson JE. Understanding the divergent data on postmenopausal hormone therapy. *N Engl J Med*. 2003;348:645–650.
10. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA*. 2002;288:321–333.
11. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
12. Flanders WD, Klein M. Properties of 2 counterfactual effect definitions of a point exposure [Erratum in: *Epidemiology* 2008;19:168]. *Epidemiology*. 2007;18:453–460.
13. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986;7:1393–1512. [Errata in *Computers and Mathematics with Applications* 1987;14:917–921. Addendum in *Computers and Mathematics with Applications* 1987;14:923–945. Errata to addendum in *Computers and Mathematics with Applications*. 1987;18:477].
14. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, et al. eds. *Advances in Longitudinal Data Analysis*. New York: Chapman and Hall/CRC Press; 2009.
15. van der Laan MJ, Petersen ML, Joffe MM. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *Int J Biostat*. 2005;1:article 4. (Electronic article). Available at: <http://www.bepress.com/ijb/vol1/iss1/4>.
16. Robins JM, Hernán MA, Rotnitzky A. Invited commentary: effect modification by time-varying covariates. *Am J Epidemiol*. 2007;166:994–1002.

Author's Response, Part II

Walter C. Willett,^a JoAnn E. Manson,^b Francine Grodstein^c

We encourage new ways of evaluating data, and for this reason the analysis led by Hernan et al of estrogen plus progestin menopausal hormone therapy¹ was a useful exercise. At the same time, we agree, however, with the comments of Stampfer² and Prentice³ regarding the simulated “intention to treat” analysis of epidemiologic data. In this application, and perhaps others, this methodology does not seem to provide insights that cannot be obtained with existing approaches, and it can add great obscurity and complexity; therefore, we do not recommend its routine use.

Regarding other comments by Prentice³ and Hoover,⁴ the inclusion by Hernan et al of cases of coronary heart disease (CHD) occurring between the onset of hormone use and the first follow-up questionnaire, which were not included in the previous, traditional strictly prospective analysis, had only a very small effect on the overall association between hormone use and CHD. We had included these cases in an earlier sensitivity analysis,⁵ and the CHD risk reduction among recently menopausal women who used hormones remained strong and statistically significant. More important, to simulate the WHI trial in the analysis by Hernan et al, the majority of CHD cases occurring within 10 years of menopause (in which hormone use was inversely related to risk of CHD) was eliminated. This gave more weight to hormone use starting more than 10 years after menopause (where hormone use was not associated with lower risk of CHD), which does not represent actual practice or the previous results from the NHS.

From these and other analyses comparing the results for CHD from the randomized trial and observational studies, it is apparent that the differences are primarily due to different distributions in time since menopause before starting hormones. Thus, as emphasized by Hoover⁴ and Mendelsohn and Karas,⁶ the overall Women's Health Initiative results for CHD should not be generalized to the majority of women who start hormone therapy near the time of menopause, and correspondingly the overall results of the NHS and other observational results should not be generalized to women starting hormone use many years after menopause. As noted by Hoover, we have learned much about the complexities surrounding time of initiating hormone use, which will need to be considered in future studies of various formulations, doses, and durations.

REFERENCES

1. Hernan MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19:766–779.
2. Stampfer MJ. ITT for observational data—worst of both worlds? *Epidemiology*. 2008;19:783–784.
3. Prentice RL. Data analysis methods and the reliability of analytic epidemiologic research. *Epidemiology*. 2008;19:785–788.
4. Hoover RN. The sound and the fury: was it all worth it? *Epidemiology*. 2008;19:780–782.
5. Grodstein F, Manson JE, Stampfer MJ. Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. *J Womens Health (Larchmt)*. 2006;15:35–44.
6. Mendelsohn ME, Karas RH. HRT and the young at heart. *N Engl J Med*. 2007;356:2639–2641.

From the ^aDepartment of Nutrition, Harvard School of Public Health, and Department of Medicine, Channing Laboratory, Brigham and Women's Hospital, Boston, Massachusetts; ^bBrigham and Women's Hospital, Division of Preventive Medicine, Boston, Massachusetts; and ^cDepartment of Medicine, Channing Laboratory, Brigham and Women's Hospital, Boston, Massachusetts.

Copyright © 2008 by Lippincott Williams & Wilkins

ISSN: 1044-3983/08/1906-0793

DOI: 10.1097/EDE.0b013e318188e84e