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

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

1. Introduction

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

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

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

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

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

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

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

   Similarly, WHI hormone users who stayed on hormone therapy for extended periods and those who discontinued hormone therapy shortly after initiation may not be comparable because treatment discontinuation was not randomized. The nonnull WHI results, however, cannot be explained by confounding for treatment discontinuation because the analysis was conducted under the intention-to-treat (ITT) principle. That is, the effect of hormone therapy was estimated by comparing the CHD risk of those randomly assigned to hormone therapy and placebo, regardless of whether they complied with their assigned treatment. The ITT effect will generally be closer to the null than the effect had all women fully complied with their assigned treatment.

3. Imprecise ascertainment of the time of hormone therapy initiation

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

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

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

2. Study Protocol of the GPRD Trials

2.1 *Eligibility Criteria*

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

2.2 *Baseline and Follow-Up*

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

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

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

2.3 *Treatment Regimes*

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

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

2.4 *Ascertainment of CHD Endpoints and Confounding Variables*

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

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

3. Analytic Approach for the GPRD Trials

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

3.1 *ITT Effect of Treatment*

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

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

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

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

3.2 *Effect of Continuous Treatment*

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

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

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

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

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

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

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

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

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

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

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

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

4. Results from the GPRD Trials

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

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

4.1 *Estimates of the ITT Effect*

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

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

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

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

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

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

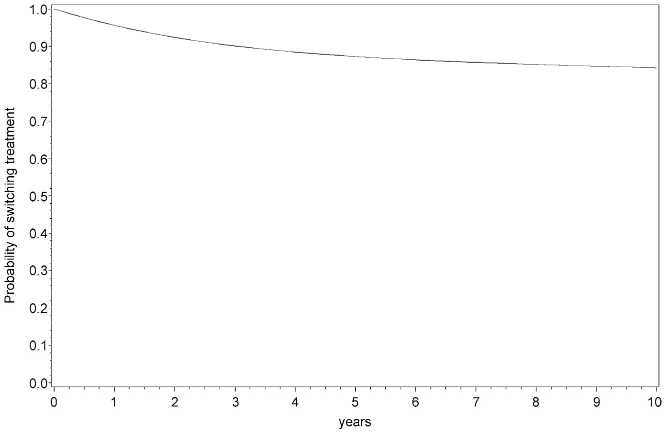

Figure 1. Probability of initiating hormone therapy among noninitiators.

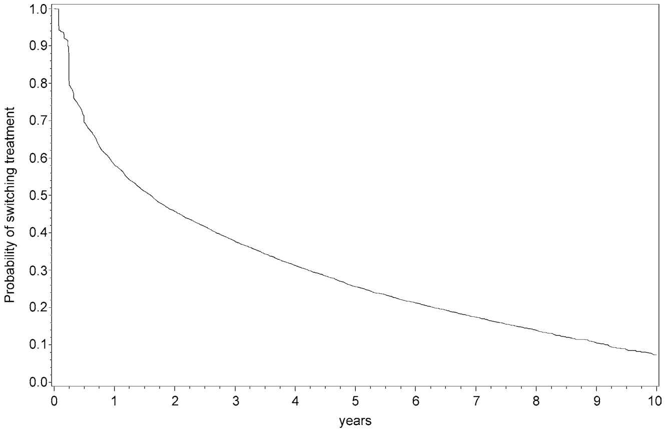

Figure 2. Probability of discontinuing hormone therapy among initiators.

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

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

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

5. Discussion and Comparison with Prentice et al.

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

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

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

Our approach provides unbiased estimates of the ITT effect only under the assumption of no unmeasured confounders for treatment initiation. Although this assumption cannot be directly tested in observational studies, comparison between the adjusted and the unadjusted estimates can be useful in assessing the hypothesis that substantial confounding by unmeasured factors remains. When we repeated our ITT analysis without adjustment for baseline covariates (except age and calendar month), the estimated hazard ratio was 0.85 (95% CI: 0.67, 1.08), which is only moderately less than the fully adjusted estimate 0.92. Were sampling variability absent, it would then follow that the magnitude of confounding due to unmeasured variables would have to exceed the confounding due to measured variables to explain the full GPRD-WHI discrepancy. Given the breadth of the measured variables, we believe this hypothesis seems unlikely, although a downward bias of perhaps 0.1 or 0.2 in our hazard ratio estimate is still plausible, especially in light of the large sampling variability. Indeed large sampling variability is a major problem. For example, the overall ITT hazard ratios from the GPRD and the WHI trials were estimated with similarly low precision (width of the 95% CIs on the log scale: about 0.46 in WHI and 0.47 in GPRD) with point estimates close to the null. This relatively low precision precludes drawing strong conclusions from either study and produces overlapping confidence intervals for the GPRD and the WHI estimates and a nonsignificant estimated difference in ITT effects. For the all-cause mortality hazard our GPRD estimates were quite similar to the WHI estimate of 0.98 (95% CI: 0.70, 1.37).

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

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

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

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

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

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

Prentice et al. mention the existence of methods for analyzing double blind randomized trial suffering from noncompliance that both (i), like an as treated analysis, provide estimates of the treatment effect under full compliance and yet (ii), like an ITT analysis, protect the α-level under the null hypothesis of no treatment effect (without imposing any assumptions concerning either the existence or magnitude of unmeasured confounding for treatment continuation). Specifically, Prentice et al. reference methodologies proposed by Cuzick, Edwards, and Segnan (1997) and Frangakis and Rubin (1999). However, these methodologies only apply if compliance is of the "all or none" type, and censoring by end of follow-up is independent conditional on complier type. But in the WHI compliance is complex and time varying with women repeatedly stopping and starting their assigned therapy. Further, although less likely, censoring by end of follow-up may be dependent if secular changes in baseline mortality risk have occurred over the trial accrual period. In this setting, as far as we are aware, *g*-estimation of nested structural models (usually referred to as SNFTMs) is the only general methodology available for the analysis of failure time data that satisfies both (i) and (ii) (Mark and Robins, 1993). Of course, adequate data on actual treatment $A(t)$ must be available for analysis. The Appendix provides further detail.

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

G-Estimation of Nested Structural Models for Survival Analysis

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

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

Duncan C. Thomas

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

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

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