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' 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. Prentice[4] and Hoover[5] are positive, whereas Stampfer[6] 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 estimates[1] although, as pointed out by Prentice[4] 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 [a]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts; [b]Departments 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."

ISSN: 1044-3983/08/1906-0789
DOI: 10.1097/EDE.0b013e318188e85f

a recent NHS paper[2] 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 nonadherent. 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 methodologic explanations[9] are that the NHS observational estimates are biased for one or more of the following reasons:

a. 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.
b. 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.
c. 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 paper[3]). 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 observatio*nal data* - *w*orst 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.