Editors' note: *This series addresses topics that affect epidemiologists across a range of specialties. Commentaries start as invited talks at symposia organized by the Editors. This paper was presented at the 2009 Society for Epidemiologic Research Annual Meeting in Anaheim, CA.*

The Hazards of Hazard Ratios

Miguel A. Hernán

The hazard ratio (HR) is the main, and often the only, effect measure reported in many epidemiologic studies. For dichotomous, non–time-varying exposures, the HR is defined as the hazard in the exposed groups divided by the hazard in the unexposed groups. For all practical purposes, hazards can be thought of as incidence rates and thus the HR can be roughly interpreted as the incidence rate ratio. The HR is commonly and conveniently estimated via a Cox proportional hazards model, which can include potential confounders as covariates.

Unfortunately, the use of the HR for causal inference is not straightforward even in the absence of unmeasured confounding, measurement error, and model misspecification. Endowing a HR with a causal interpretation is risky for 2 key reasons: the HR may change over time, and the HR has a built-in selection bias. Here I review these 2 problems and some proposed solutions. As an example, I will use the findings from a Women's Health Initiative randomized experiment that compared the risk of coronary heart disease of women assigned to combined (estrogen plus progestin) hormone therapy with that of women assigned to placebo.[1] By using a randomized experiment as an example, the discussion can focus on the shortcomings of the HR, setting aside issues of confounding and other serious problems that arise in observational studies.

The Women's Health Initiative followed over 16,000 women for an average of 5.2 years before the study was halted due to safety concerns. The primary result from the trial was a HR. As stated in the abstract[1] and shown in Table 1 of the article, "Combined hormone therapy was associated with a hazard ratio of 1.24."[1] In addition, Table 2 provided the HRs during each year of follow-up: 1.81, 1.34, 1.27, 1.25, 1.45, and 0.70 for years 1, 2, 3, 4, 5, and 6+, respectively. Thus, the HR reported in the abstract and Table 1 can be viewed as some sort of weighted average of the period-specific HRs reported in Table 2.

This bring us to Problem 1: although the HR may change over time, some studies report only a single HR averaged over the duration of the study's follow-up. As a result, the conclusions from the study may critically depend on the duration of the follow-up. For example, the average HR in the WHI would have been 1.8 if the study had been halted after 1 year of follow-up, 1.7 after 2 years,[2] 1.2 after 5 years, and—who knows—perhaps 1.0 after 10 years. The 24% increase in the rate of coronary heart disease that many researchers and journalists consider as *the* effect of combined hormone therapy is the result of the arbitrary choice of an average follow-up period of 5.2 years. A trial with a shorter follow-up could have reported an 80% increase, whereas a longer trial might have found little or no increase at all.

From the Department of Epidemiology, Harvard School of Public Health, and the Harvard-MIT Division of Health Sciences and Technology, Boston, MA.
Supported by funds from NIH grant R01 HL080644.
Editors' note: *Related articles appear on pages 10 and 3.*
Correspondence: Miguel A. Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. E-mail: miguel_hernan@post.harvard.edu.

ISSN: 1044-3983/10/2101-0013
DOI: 10.1097/EDE.0b013e3181c1ea43

The magnitude of the average HR depends on the length of follow-up because the average HR ignores the distribution of events during the follow-up. The average HR can take the value 1.0 if the hazard in the exposed is identical to the hazard in the unexposed during the entire follow-up, or if the hazard in the exposed is higher during, say, the first 5 years and lower afterward. Incidentally, the same problem arises whether the average HR is directly estimated in a cohort study, as discussed here, or estimated via the odds ratio of a properly designed case-control study with incidence density sampling.

One might then conclude that we should forget about the average HR and restrict our attention to the period-specific HRs, which seem to capture the potentially time-varying magnitude of the effect. This brings us to Problem 2: the period-specific HRs have a built-in selection bias. To describe the bias, consider that the (discrete-time) hazard during period t is defined as the risk of the outcome during period t among those who reached period t free of the outcome. In the Women's Health Initiative, the calculation of the HR during year t was restricted to women who did not develop coronary heart disease—the "survivors"—between baseline and the beginning of year t. The HR after year 5 was 0.7, which means that the disease rate after year 5 was lower in the treatment arm (the hazard in the numerator of the HR) than in the placebo arm (the hazard in the denominator).

However, this apparently protective effect of hormone therapy after year 5 is hardly surprising if one bears in mind that women vary in their susceptibility to heart disease. A certain proportion of all women enrolled in the trial were particularly prone to develop heart disease if they were exposed to hormone therapy or other factors (for simplicity, let's refer to them as the "susceptible women"). The proportion of susceptible women in the trial was of course unknown but, because of randomization, it was expected to be the same in both the treatment and placebo arms at baseline. However, these susceptible women were preferentially excluded from the treatment arm as they developed heart disease over time—precisely because they were assigned to a therapy with harmful effects to which they were susceptible (all other factors to which they were susceptible were expected to be equally distributed between the 2 arms). With time, the proportion of susceptible women progressively increased in the placebo arm compared with the treatment arm. The bias due to the differential selection of less susceptible women over time, because of differential depletion of susceptibles, is the built-in selection bias of period-specific HRs. This bias may explain that the HR after year 5 is less than 1.0 even if hormone therapy has no truly preventive effect in any woman at any time. This built-in selection bias of the HR has also been described using causal diagrams.[3,4]

In short, the average HR may be uninformative because of potentially time-varying period-specific HRs, and because the period-specific HRs may be time-varying because of built-in selection bias. These problems can be overcome by summarizing the study findings as appropriately adjusted survival curves, where the survival at time t is defined as the proportion of individuals who are free of disease through time t. Another alternative not discussed here is the comparison of the distribution of survival times between the exposed and the unexposed, which can be accomplished by using accelerated failure time models[5] rather than Cox models.

Because of the shortcomings of the HR, the analysis of randomized experiments routinely include Kaplan-Meier survival curves—or their complement, the cumulative risk curve (see Figure 2 of the Women's Health Initiative trial report[1]). In contrast (and despite multiple warnings in the epidemiologic literature[3–6]), the analysis of observational follow-up studies are commonly summarized by HRs only. A possible explanation for this practice in observational studies is the need to deal with confounding.

The HRs presented in observational studies are not simply the hazard in the exposed divided by the hazard in the unexposed. Rather, these HRs are adjusted for measured confounders by using regression models, inverse probability weighting, or other methods. Unadjusted HRs would be of little use for causal inference from observational data, as would unadjusted survival curves. It is not unexpected that most epidemiologic articles include HRs only, because epidemiology students are traditionally taught to estimate adjusted HRs but not adjusted survival curves.[7] The next paragraph sketches a general procedure to obtain survival curves adjusted for baseline confounders.

First, fit a discrete-time hazards model (eg, a pooled logistic model with relatively short periods) that estimates, at each time and for each person, the conditional probability of remaining free of the outcome given exposure, baseline covariates, and time of follow-up. Allow for time-varying hazards by modeling the variable "time of follow-up," using a flexible functional form (eg, cubic splines), and for time-varying HRs by adding product terms between exposure and "time of follow-up." Second, for each subject, multiply the model's predicted values through time t to estimate the survival at t for subjects with their same combination of covariate values. One can then construct conditional (adjusted) survival curves under the conditions of exposure and no exposure for each observed combination of values of the baseline covariates (in randomized trials, the survival curves are unconditional or marginal, ie, averaged over all the individuals irrespective of their covariate values). Third, predict the survival at time t for each subject both under exposure and under no exposure, regardless of the subject's exposure status. Fourth, separately average the conditional survivals under exposure and under no exposure, over all subjects. This last step effectively standardizes the curves to the empirical distribution of the covariates in the study, and

results in 2 marginal survival curves: one under exposure, another under no exposure.

The above procedure can be extended in a number of ways. In settings with time-varying exposures and confounders, the procedure can be combined with inverse probability weighting of the hazards model. This procedure has been used to present adjusted survival curves under continuous use ("always exposed") and no use of hormone therapy ("never exposed") in the analysis of both observational studies[8] and randomized experiments[9] in which time-varying exposures arise when considering adherence-adjusted analyses. In settings with continuous rather than dichotomous exposures, the procedure requires the choice of a finite number of levels of exposure to be compared ("always versus never exposed" will not do).[10] One may then construct as many survival curves as there are exposure levels of interest. For continuous and time-varying exposures one needs to be especially careful about dose-response assumptions. Sensitivity analyses can be used to evaluate the possibility of model extrapolation beyond the observed data. Confidence intervals for the survival curves can be obtained by bootstrapping.

So should we outlaw the use of HRs in epidemiologic studies? Of course not. A single average HR through t may be misleading, as explained above, but a single survival probability at t could be as misleading because both measures ignore the distribution of events between baseline and t. On the other hand, a series of average HRs for increasingly longer periods of follow-up is informative. For example, in the WHI the average HRs for 1, 2, and 5 years were approximately 1.8, 1.7, and 1.2, which indicates that hormone therapy increases the cumulative risk of heart disease in the early part of the follow-up but probably not much over longer periods. The same conclusion is drawn from the survival curves for the treatment and placebo groups, which converge after 8 years. In mortality studies with sufficiently long follow-up, the survival probabilities in both groups are ensured to reach the value 0, and the average HR is ensured to reach the value 1.

An advantage of the survival curves over a series of average HRs is that the survival curves provide information about the absolute risks. For example, in the Women's Health Initiative, the average HR of 1.8 during year 1 means that the one-year risk was about 0.49% in the treatment group and 0.28% in the placebo group rather than, say, 49% versus 28%. An advantage of the average HRs over the survival curves is the readiness with which confidence intervals can be computed in standard software.

What about period-specific HRs? Their built-in selection bias makes them difficult to interpret as a measure of time-varying effect. For example, in the Women's Health Initiative, the HR goes from greater than 1.0 to less than 1.0 after year 5—that is, the hazards of the treatment and the placebo groups cross at about year 5. However, this crossing of hazards is essentially meaningless from a practical standpoint. What really matters is that the survival is lower in the placebo group compared with the treatment group until at least year 8. Hazards may cross at some point during the follow-up because of depletion of susceptibles even if the survival curves never cross. Cumulative measures, such as a series of average HRs or survival curves, are needed to summarize the data in a meaningful way. On the other hand, period-specific HRs are useful as an intermediate step to estimate survival curves in the procedure described above.

In summary, survival curves are more informative than HRs and can be easily generated. It would not be a bad thing to see them more widely used in observational studies.

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. Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. *Biometrics*. 2005;61:899–911.
3. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
4. Flanders WD, Klein M. Properties of 2 counterfactual effect definitions of a point exposure. *Epidemiology*. 2007;18:453–460.
5. Hernán MA, Cole SR, Margolick JB, Cohen MH, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol Drug Saf*. 2005;14:477–491.
6. Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*. 1996;7:498–501.
7. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*. 2004;75:45–49.
8. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease (with discussion). *Epidemiology*. 2008;19:766–779.
9. Toh SG, Hernández-Díaz S, Logan R, Rossouw JE, Hernán MA. Coronary heart disease in postmenopausal users of estrogen plus progestin hormone therapy: Does the increased risk ever disappear? *Ann Intern Med.* In Press.
10. Zhang Y, Thamer M, Cotter D, Kaufman J, Hernán MA. Estimated effect of epoetin dosage on survival among elderly hemodialysis patients in the United States. *Clin J Am Soc Nephrol*. 2009;4:638–644.