American Journal of Epidemiology
Vol. 168, No. 6
DOI: 10.1093/aje/kwn164
Advance Access publication August 5, 2008

Practice of Epidemiology

Constructing Inverse Probability Weights for Marginal Structural Models

Stephen R. Cole[1] and Miguel A. Hernán[2,3]

[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.
[2] Department of Epidemiology, Harvard School of Public Health, Boston, MA.
[3] Harvard–MIT Division of Health Sciences and Technology, Cambridge, MA.

Received for publication January 22, 2008; accepted for publication May 12, 2008.

The method of inverse probability weighting (henceforth, weighting) can be used to adjust for measured confounding and selection bias under the four assumptions of consistency, exchangeability, positivity, and no misspecification of the model used to estimate weights. In recent years, several published estimates of the effect of time-varying exposures have been based on weighted estimation of the parameters of marginal structural models because, unlike standard statistical methods, weighting can appropriately adjust for measured time-varying confounders affected by prior exposure. As an example, the authors describe the last three assumptions using the change in viral load due to initiation of antiretroviral therapy among 918 human immunodeficiency virus-infected US men and women followed for a median of 5.8 years between 1996 and 2005. The authors describe possible tradeoffs that an epidemiologist may encounter when attempting to make inferences. For instance, a tradeoff between bias and precision is illustrated as a function of the extent to which confounding is controlled. Weight truncation is presented as an informal and easily implemented method to deal with these tradeoffs. Inverse probability weighting provides a powerful methodological tool that may uncover causal effects of exposures that are otherwise obscured. However, as with all methods, diagnostics and sensitivity analyses are essential for proper use.

bias (epidemiology); causality; confounding factors (epidemiology); probability weighting; regression model

Abbreviations: AIDS, acquired immunodeficiency syndrome; HAART, highly active antiretroviral therapy; HIV, human immunodeficiency virus; HIV-1, human immunodeficiency virus type 1.

Inverse probability weighting (henceforth, weighting) can be used to estimate exposure effects. Unlike standard statistical methods, weighting can appropriately adjust for confounding and selection bias due to measured time-varying covariates affected by prior exposure (1).

Under the four assumptions of consistency, exchangeability, positivity, and no misspecification of the model used to estimate the weights, weighting creates a pseudo-population in which the exposure is independent of the *measured* confounders (2). The pseudo-population is the result of assigning to each participant a weight that is, informally, proportional to the participant's probability of receiving her own exposure history. In such a pseudo-population, one can regress the outcome on the exposure using a conventional regression model that does not include the measured confounders as covariates. Fitting a model in the pseudo-population is equivalent to fitting a weighted model in the study population. The parameters of such weighted regression models, which equal the parameters of marginal structural models (3), can be used to estimate the average causal effect of exposure in the original study population.

In recent years, several published estimates of the effect of time-varying exposures have been based on weighted estimation of the parameters of marginal structural models (4–24). Most of these articles discuss the plausibility of the exchangeability assumption, often referred to as the

Correspondence to Dr. Stephen R. Cole, Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, McGavran-Greenberg Hall, Campus Box 7435, Chapel Hill, NC 27599-7435 (e-mail: cole@unc.edu) (present address).

assumption of no unmeasured confounding, and emphasize correctly that this assumption is not empirically verifiable. These articles also implicitly assume that consistency holds, which is a reasonable assumption when estimating the effect of medical treatments (refer to appendix 2 for a formal definition of consistency). However, these articles do not usually include an explicit discussion of the role of the other three assumptions stated above. Here, we describe the role of three of the four assumptions in weighted estimation and the interpretation of results. This paper is structured as follows. First, we describe a motivating example from our ongoing work in human immunodeficiency virus (HIV) epidemiology. Second, in the context of our motivating example, we describe the assumptions of exchangeability, positivity, and no model misspecification, as well as the tradeoffs that an epidemiologist may encounter when attempting to make inferences under these assumptions. Third, we describe an informal method to deal with these tradeoffs. We conclude with a brief discussion and some recommendations for constructing inverse probability weights (henceforth, weights).

EXAMPLE: ANTIRETROVIRAL THERAPY AND VIRAL LOAD IN HIV-INFECTED INDIVIDUALS

To provide motivation for our discussion, we use the analysis reported in a recent paper (19) that estimated the effect of initiation of highly active antiretroviral therapies (HAART) on the change in human immunodeficiency virus type 1 (HIV-1) RNA viral load in HIV-infected individuals. We suggest reading reference 19 in concert with the present paper. In brief, 918 HIV-infected men and women not using HAART at study entry were seen semiannually in the Multicenter AIDS [acquired immunodeficiency syndrome] Cohort Study or Women's Interagency HIV Study for a median of 5.8 years between 1996 and 2005. We estimated the effect of time-varying HAART initiation on change in $\log_{10}$ viral load.

For each subject i and visit j, we estimated a weight SW_{ij} that was, informally, proportional to the inverse (or reciprocal) of the probability of having her own observed exposure and censoring history through that visit. For a formal definition of the weights, refer to appendix 1. We then fit a weighted repeated measures regression model in which an individual was assigned her estimated weight SW_{ij} at each visit. The primary effect estimate was an immediate and sustained 1.91 $\log_{10}$ decrease in viral load after HAART initiation. Next, we describe the assumptions necessary for valid inferences and their practical implications in the context of our example.

EXCHANGEABILITY

Exchangeability implies the well-known assumption of no unmeasured confounding. For the assumption of no unmeasured confounding to hold, we have to measure enough joint predictors of exposure and outcome such that, within the levels of these predictors, associations between exposure and outcome that are due to their common causes will disappear. For a formal definition, refer to appendix 2.

Exchangeability assumptions are not testable in observed data, but one may explore the sensitivity of inferences from weighted regression to this assumption by using sensitivity analysis as described by Robins (25) and implemented by various authors (10, 14, 26). We do not reiterate the approach to sensitivity analysis here but, rather, assume that the most important confounders were identified using expert knowledge (27, 28) and were then appropriately measured and included in the analysis. Specifically, we assumed that conditioning on several baseline covariates and the most recent values of CD4 cell count and viral load is sufficient to achieve exchangeability between those who did and did not initiate therapy at any time during the follow-up. Later, in table 3, we assess the impact of adding further potential confounders. As a consequence of our assumption that therapy is continuously used after initiation, we do not need to assume that those who did and did not *discontinue* were exchangeable, and hence our estimates do not require the assumption of exchangeability after therapy initiation. The price we pay for this intent-to-treat assumption is, of course, some bias toward the null that increases with the number of participants that discontinue therapy during follow-up.

As a practical rule to help ensure approximate exchangeability, it may appear obvious that investigators need first to identify and measure as many potential confounders as possible. Then, investigators would include those potential confounders in the model used to estimate the denominator of the weights. However, this strategy may not always decrease net bias in finite samples for two reasons. First, the addition of a nonconfounding variable may *introduce* selection bias due to collider stratification (29, 30). Second, adding too many potential confounders in relation to the number of observations may introduce finite-sample bias, which is related to the bias due to nonpositivity discussed below. Further, adding nonconfounding variables to the model for the weights may decrease the statistical efficiency of the effect estimate (i.e., yield wider confidence intervals) (31). For these reasons and as illustrated below, in practice one may not always want to include as many potential confounders as possible.

POSITIVITY

For any method that estimates the average causal effect in the study population, one must be able to estimate the average causal effect in each subset of the population defined by the confounders. For example, to estimate the effect of HAART in the presence of confounding by CD4 cell count, we need to be able to estimate the effect of HAART in every category of CD4 cell count. An effect cannot be estimated in a subset of the study population if *everyone* is exposed (or unexposed) in that subset. Positivity is the condition that there are both exposed and unexposed individuals at every level of the confounders. For a formal definition, refer to appendix 2. Positivity is guaranteed (unconditionally) in experiments because, by design, there will be individuals assigned to each level of the studied treatment. The positivity

assumption is also called the experimental treatment assumption (32). Because the weights SW_{ij} can always be estimated parametrically from the data, even in the presence of violations of the positivity assumption, lack of positivity (like lack of consistency) may go undetected unless explicitly investigated.

If somebody cannot *possibly* be exposed at one or more levels of the confounders, then positivity is violated because there is a structural zero probability of receiving the exposure. To fix this idea, we provide two examples. First, in an occupational epidemiology study to estimate the health effects of a certain chemical, being at work is a potential confounder often used as a proxy for health status. If one cannot be exposed to the chemical outside the workplace, then there is a structural zero probability of exposure to the chemical among those no longer at work. Second, in a pharmacoepidemiology study to estimate the effects of a particular drug, an absolute contraindication for treatment (e.g., liver disease) may be a surrogate for bad prognosis. If one cannot possibly be treated in the presence of the contraindication, then there is a structural zero probability of receiving the treatment among those with the contraindication. An obvious solution is restricting the inference to the subset with a positive probability of exposure. However, if the structural zero occurs within levels of a time-varying confounder (e.g., liver disease), then restriction or censoring may lead to bias, whether one uses weighting or other methods (30).

Even in the absence of structural zeros, random zeros (also called practical violations of the experimental treatment assumption (33)) may occur by chance because of small sample sizes or high dimensional (i.e., highly stratified or continuous) data. Even a relatively large study may have zero proportions for particular exposure and covariate histories as the number of covariates increases. In fact, when modeling continuously distributed covariates, random zeros are essentially guaranteed because of the infinite number of possible values. In such cases, the use of parametric models smoothes over the random zeros by borrowing information from individuals with histories similar to those that, by chance, resulted in random zeros. For example, in table 1 we present the proportions of HAART initiation (i.e., exposure) at 25 levels of joint time-varying CD4 cell count and viral load. At two of 25 levels, we see nonpositivity or a zero proportion exposed. These observed zero proportions may be structural or random. In table 1, both zero proportions occur in person-time contributions where immunity is not depleted (i.e., CD4 count, >749 cells/mm^3) but virus is detectable. On the basis of prior substantive knowledge and surrounding nonzero proportions, we concluded that these two nonpositive proportions appear to be random zeros, rather than structural zeros, and thus proceeded to model the probability of exposure to construct weights.

There is a tradeoff between reducing confounding bias and increasing bias and variance due to nonpositivity. Data become sparse, and the likelihood of random zeros (and hence bias due to nonpositivity) increases as one includes more confounders. For example, in table 2, we progressively expand the number of categories used to define CD4 count and viral load in the construction of weights from one to nine categories. Table 2 also presents the effect estimate (i.e., the difference in $\log_{10}$ viral load) and its standard error obtained by bootstrap. Estimated weights with the mean far from one or very extreme values are indicative of nonpositivity or misspecification of the weight model, and thus table 2 also presents the mean, standard deviation, minimum, and maximum estimated weights. As the number of categories increases from one to five, we observe three changes. First, the effect estimate increases in absolute value, which (in the present substantive setting) suggests a better control of confounding. Second, the precision of the effect estimate decreases. Third, the standard deviation and range of the weights increase, which is the cause of the decreasing precision of the effect estimate. For seven categories of CD4 cell count and viral load, the effect estimate moves toward the null and its standard error triples. For nine categories of CD4 cell count and viral load, the weights become so alarmingly variable (with a mean no longer equal to one) that the effect estimate is no longer computable.

TABLE 1. Proportions of 286 HAART* initiators observed in 4,778 semiannual study visits by categories of prior time-varying CD4 and HIV-1* RNA viral load, Multicenter AIDS* Cohort Study and Women's Interagency HIV* Study, 1996–2005

CD4 count, cells/mm^3	Viral load, copies/ml	No. exposed	No. of person-visits	Proportion
>749	<401	2	308	<0.01
	401–<4,000	3	253	0.01
	4,000–10,000	0	278	0
	10,001–35,000	4	117	0.03
	>35,000	0	38	0
501–749	<401	3	199	0.02
	401–<4,000	3	354	<0.01
	4,000–10,000	5	374	0.01
	10,001–35,000	15	259	0.06
	>35,000	14	162	0.09
351–500	<401	2	76	0.03
	401–<4,000	12	268	0.04
	4,000–10,000	5	263	0.02
	10,001–35,000	25	280	0.09
	>35,000	17	247	0.07
200–350	<401	1	36	0.03
	401–<4,000	5	118	0.04
	4,000–10,000	6	162	0.04
	10,001–35,000	17	242	0.07
	>35,000	55	273	0.20
<200	<401	3	12	0.25
	401–<4,000	3	25	0.12
	4,000–10,000	4	53	0.08
	10,001–35,000	13	101	0.13
	>35,000	69	280	0.25
Total		286	4,778	

* HAART, highly active antiretroviral therapy; HIV-1, human immunodeficiency virus type 1; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus.

TABLE 2. Effect of HAART* versus no HAART on change in HIV-1* RNA viral load under a series of models using increasingly fine categorization of time-varying CD4 count and viral load in construction of inverse probability weights, Multicenter AIDS* Cohort Study and Women's Interagency HIV* Study, 1996–2005†

No. of categories‡	Estimated weights		Difference in viral load, $\log_{10}$ copies/ml	
	Mean (SD*)	Minimum/ maximum	Estimate	SE*,§
1	1.00 (0)	1.00/1.00	−1.59	0.089
3	1.01 (0.96)	0.15/33.5	−1.73	0.103
5	1.00 (1.42)	0.06/59.1	−1.79	0.125
7	1.03 (1.61)	0.06/74.2	−1.74	0.392
9	536.7 (8,037.3)	0.05/1.6 × 100,000	—¶	—

* HAART, highly active antiretroviral therapy; HIV-1, human immunodeficiency virus type 1; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; SD, standard deviation; SE, standard error.

† All models in this table stabilized weights by using only a three-knot spline for time.

‡ The nine categories of CD4 count and viral load were as follows: ≤25, 26–50, 51–100, 101–150, 151–<200, 200–350, 351–500, 501–749, >749 cells/mm^3 and ≤100, 101–1,000, 1,001–10,000, 10,001–50,000, 50,001–100,000, 100,001–200,000, 200,001–300,000, 300,001–500,000, >500,000 copies/ml, respectively; coarsened categories were obtained by collapsing adjacent outer categories.

§ The standard deviation of 500 nonparametric bootstrap sample estimates; 500, 500, 500, and 496 converged.

¶ —, not computable.

Contrary to the naïve belief that more finely defined confounders will always lead to better confounding control, table 2 shows that bias and variance of the effect estimate may increase with the number of categories. Similarly, one may wish to omit control for weak confounders that cause severe nonpositivity bias because of a strong association with exposure. In addition, although not illustrated in table 2, the magnitude of nonpositivity bias typically increases with the number of time points and decreases with the use of appropriately stabilized weights.

Weighted estimates are more sensitive to random zeros than is standard regression or stratification estimates, which implicitly extrapolate to levels of the covariates with a lack of positivity. Users of weighted approaches need tools to handle this bias-variance tradeoff. Wang et al. (33) have proposed a computationally demanding diagnostic tool to quantify the finite-sample bias due to random zeros in weighted estimates. After reviewing the assumption of no model misspecification in the next section, we propose an informal method to evaluate this bias-variance tradeoff. Refer to references 32–34 for more formal methods.

CORRECT MODEL SPECIFICATION

Weighted estimation of the parameters of marginal structural models requires fitting several models: 1) the structural (i.e., weighted) model, 2) the exposure model, and 3) the censoring model. For simplicity and because this paper focuses on constructing weights to estimate the parameters of any marginal structural model through weighted regression, we will assume throughout that the structural model is correctly specified. In practice, investigators will want to explore the sensitivity of their estimates to different structural model specifications (e.g., linear vs. threshold dose-response, long- vs. short-term effects, and so on).

To construct appropriate weights, investigators need to correctly specify the models for exposure and censoring. Here, we will discuss only modeling of the exposure distribution, but our comments apply equally to modeling the censoring distribution. As stated above, a necessary condition for correct model specification is that the stabilized weights have a mean of one (2). In table 3, we provide a step-by-step example of building weights for the marginal structural model detailed previously (19) and described above. Although the step-by-step process is a simplified representation of the actual process, we hope that sharing the general approach may guide future implementations of marginal structural models.

In specification 1, the model to estimate the denominator of the weights was a pooled logistic model for the probability of exposure initiation at each visit. Specifically, each person-visit was treated as an observation, and the model was fit on those person-visits for which no exposure had occurred through the prior visit. The covariates were linear terms for follow-up time, baseline CD4 cell count and viral load, and time-varying CD4 cell count and viral load measured at the prior visit. This model, which is a parametric discrete-time approximation of the Cox proportional hazards model for exposure initiation (35, 36), assumes that the relation between the baseline covariates (and follow-up time) and the probability of exposure initiation is linear on the logit scale. The model to estimate the numerator of the weights was also a pooled logistic model for the probability of exposure initiation, except that time-varying CD4 cell count and viral load were not included as covariates. The mean of the estimated weights was 1.07 (standard deviation, 1.47), the 1/minimum and maximum estimated weights were 33.3 and 26.4, and the effect estimate was −1.94 (standard error, 0.17).

In specification 2, we replace the linear terms for baseline and time-varying CD4 and viral load with categories (i.e., CD4: <200, 200–500, >500 cells/mm^3; and viral load detectable (at 400 copies/ml) or not) to illustrate the impact of potential residual confounding within categories of the confounders. The estimated weights appear better behaved than in specification 1 (e.g., the mean moves from 1.07 to 1.05, 1/minimum and maximum notably smaller), and the standard error for the difference in $\log_{10}$ viral load is a striking 39 percent ($1 - 0.104/0.170 = 0.388$) smaller, but the effect estimate of −1.66 moved closer to the unadjusted value of −1.56 (i.e., one category, table 2).

In specification 3, the numerator and denominator are as in specification 1, but we add three-knot restricted cubic splines to all linear terms. Other smoothing techniques could be used (37). This flexible parameterization of the time-varying confounders is generally preferred, because

TABLE 3. Effect of HAART* versus no HAART on change in HIV-1* RNA viral load under a series of models for the construction of inverse probability weights, Multicenter AIDS* Cohort Study and Women's Interagency HIV* Study, 1996–2005

Specification	Description	Estimated weights		Difference in viral load, $\log_{10}$ copies/ml	
		Mean (SD*)	Minimum/maximum	Estimate	SE*,†
1	Numerator includes linear terms for baseline CD4, RNA, and time. Denominator includes linear terms for baseline CD4, RNA, time, $CD4_{-1}$, and RNA_{-1}.	1.07 (1.47)	0.03/26.4	−1.94	0.170
2	Numerator and denominator are as in step 1 but replace linear terms for baseline and time-varying CD4 and RNA with step functions (i.e., categories‡).	1.05 (0.65)	0.11/16.6	−1.66	0.104
3	Numerator and denominator are as in step 1, with three-knot splines to all linear terms.	1.05 (1.17)	0.03/37.0	−1.91	0.139
4	Numerator and denominator are as in step 3, plus a product between $CD4_{-1}$ and time in denominator.	1.04 (1.15)	0.03/46.8	−1.91	0.132
5	Numerator and denominator are as in step 4, plus three-knot splines for $CD4_{-2}$ and RNA_{-2} in denominator.	1.04 (1.67)	0.03/83.5	−1.95	0.133
6	Numerator and denominator are as in step 4, plus indicators for $AIDS_{-1}$ and presence of HIV $symptoms_{-1}$ in denominator.	1.05 (1.37)	0.03/68.4	−1.91	0.130

* HAART, highly active antiretroviral therapy; HIV-1, human immunodeficiency virus type 1; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; SD, standard deviation; SE, standard error.

† The standard deviation of 500 nonparametric bootstrap sample estimates; 500 always converged.

‡ The CD4 count categories were as follows: <200, 200–500, >500 cells/mm^3; and viral load was detectable at 400 copies/ml or not.

it liberates one from much of the residual confounding or finite-sample bias inherent in categorical variables (e.g., specification 2) and reduces the potential bias due to model misspecification from strong linearity assumptions (e.g., specification 1). Compared with specification 1, the estimated weights and effect estimate are similar, but the standard error is reduced by 18 percent ($1 - 0.139/0.170 = 0.182$).

In specification 4, we added a product term between time-varying CD4 count and follow-up time suggested by clinical colleagues, which had $p = 0.03$. Compared with specification 3, there is little change in the estimated weights (although the maximum weight increases), and the effect estimate remains unaltered, but its standard error is reduced by 5 percent. This is essentially the model specification used previously (19); however, the (conservative) robust standard error reported (19) was 0.135, while the bootstrap standard error reported here is 0.132.

In specification 5, we explored more fully detailed covariate histories, using time-varying CD4 count and viral load measured two visits prior to the visit at-risk for HAART initiation in addition to values measured one visit prior. Beyond an increase in the maximum weight, no notable changes are apparent.

In specification 6, we explored the addition of two more possible time-varying confounders, namely, clinical AIDS status and HIV-related symptoms (i.e., reports of persistent fever, diarrhea, night sweats, or weight loss) at the visit prior to the visit at-risk for HAART initiation. Again, no notable changes are apparent.

WEIGHT TRUNCATION AS A MEANS TO TRADEOFF BIAS AND VARIANCE

The process discussed above and presented in table 3 illustrates how the choice of the model used to construct weights may impact the results of a marginal structural model. Our decision to settle on specification 4 of table 3 was an informal bias-variance tradeoff between the inclusion of a sufficient number of flexibly modeled confounders in the weight model and the construction of well-behaved weights (mean = 1, small range) that led to a small variance of the effect estimate. Thus, compared with the model in specification 4, models that included only linear terms for the time-varying confounders (i.e., specification 1), omitted product terms (i.e., specification 3), or included additional potential confounders (i.e., specifications 5 and 6) typically resulted in similar effect estimates with a slightly greater variance or greater model complexity. On the other hand, transforming the continuous confounders into categorical variables (i.e., specification 2) resulted in a smaller variance but probably also in insufficient confounding adjustment, as the effect estimate moved considerably toward the unadjusted

TABLE 4. Effect of HAART* versus no HAART on change in HIV-1* RNA viral load under progressive truncation of inverse probability weights, Multicenter AIDS* Cohort Study and Women's Interagency HIV* Study, 1996–2005

Truncation percentiles	Estimated weights		Difference in viral load, $\log_{10}$ copies/ml	
	Mean (SD*)	Minimum/ maximum	Estimate	SE*,†
0, 100‡	1.04 (1.15)	0.03/46.8	−1.91	0.132
1, 99	1.00 (0.58)	0.20/4.49	−1.80	0.122
5, 95	0.95 (0.36)	0.36/1.93	−1.73	0.106
10, 90	0.92 (0.27)	0.49/1.42	−1.69	0.101
25, 75	0.91 (0.12)	0.75/1.03	−1.63	0.091
50, 50‡	0.95 (0.00)	0.95/0.95	−1.59	0.089

* HAART, highly active antiretroviral therapy; HIV-1, human immunodeficiency virus type 1; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; SD, standard deviation; SE, standard error.

† The standard deviation of 500 nonparametric bootstrap sample estimates; 500 always converged.

‡ No truncation of weights corresponds to a standard marginal structural model, while setting all weights to the constant 50th percentile corresponds to the baseline adjusted model.

result. Note that the best behaved weights (by the measures of mean and small range) would simply be a constant one. However, such weights would *completely* fail to control for time-varying confounding.

One simple way to explore this bias-variance tradeoff is to progressively truncate (38) the weights as shown in table 4. Specifically, the weights are progressively truncated by resetting the value of weights greater (lower) than percentile p (100 – p) to the value of percentiles p (100 – p). The first row in table 4 corresponds to the standard marginal structural model (i.e., specification 4 in table 3), while the last row in table 4 corresponds to a baseline-adjusted model (i.e., one category, table 2, or reference 19, p. 222). Assuming that the marginal structural model estimate is correct, one can see the growing bias as the weights are progressively truncated. Simultaneously, one can see the increasing precision as the weights are progressively truncated. In this case and under the assumption that the marginal structural model estimate is unbiased, the small increase in precision due to weight truncation is outweighed by the relatively large bias induced. However, here, one could reasonably argue in favor of reporting the result with the weights truncated at the first and 99th percentiles, on the basis of the centering of the weights at one and the order of magnitude reduction in the 1/minimum and maximum weights.

The requirement of a mean of one applies to the estimated weights at each time point, but, as a simplification, we pooled the estimated weights from all time points in the study. It is therefore logically possible that the chosen weight model results in a mean estimated weight closer to one than an alternative weight model but that the chosen weight model is badly misspecified for some time points, whereas the alternative weight model is slightly misspecified at all time points. Depending on the aims of analysis, we may prefer the alternative weight model over the chosen.

CONCLUSION

The construction of inverse probability weights for marginal structural models (4–20, 22), or other uses (30, 39), requires a thoughtful process including determination of a proper set of covariates upon which one can tolerate the assumptions of no unmeasured confounding and no informative censoring, exploration of positivity, and determination of a model specification that optimizes bias reduction and precision. Nonweighting methods are also subject to these same assumptions. Indeed, a process similar to that laid out here should be undertaken in *any* observational data analysis. Here, we detailed some approaches to the construction of such weights using an example from a recently published paper.

Future research is needed to formally compare competing methods to balance bias and variance when selecting from potential confounders and functional forms. In the meantime, we recommend the following: 1) Check positivity for important confounders as illustrated in tables 1 and 2. 2) Explore exchangeability by using a variety of potential confounders and functional forms as illustrated in table 3, coupled with sensitivity analysis (14). 3) Check weight model misspecifications by exploring the distribution of weights. The tradeoffs implied by the need to simultaneously guarantee exchangeability, positivity, and no model misspecifications can be explored by evaluating the sensitivity of inferences to truncating extreme weights as illustrated in table 4. In manuscripts, we encourage both acknowledging the sensitivity of the effect estimates to the weight model specification and reporting an effect estimate that is robust to different weight model specifications. Often, this will mean selecting as the main finding an effect estimate that is less extreme than that produced by certain weight model specifications. Inverse probability weighting provides a powerful methodological tool that may uncover causal effects of exposures that are otherwise obscured, but powerful tools can be dangerous if not handled with care.

ACKNOWLEDGMENTS

Dr. Cole was supported in part by National Institute of Allergy and Infectious Diseases grant R03-AI071763, and Dr. Hernán was supported in part by National Institute of Allergy and Infectious Diseases grant R01-AI073127. The Multicenter AIDS Cohort Study is funded by the National Institute of Allergy and Infectious Diseases, with additional supplemental funding from the National Cancer Institute (grants UO1-AI-35042, 5-MO1-RR-00722 (General Clinical Research Center), UO1-AI-35043, UO1-AI-37984, UO1-AI-35039, UO1-AI-35040, UO1-AI-37613, and UO1-AI-35041). The Women's Interagency HIV Study is also funded by the National Institute of Allergy and Infectious

Diseases, with supplemental funding from the National Cancer Institute, the National Institute of Child Health and Human Development, the National Institute on Drug Abuse, and the National Institute of Craniofacial and Dental Research (grants U01-AI-35004, U01-AI-31834, U01-AI-34994, AI-34989, U01-HD-32632 (National Institute of Child Health and Human Development), U01-AI-34993, and U01-AI-42590).

The authors thank Drs. Maya Petersen, Lisa Bodnar, Sander Greenland, Jonathan Sterne, and James Robins for expert advice.

Data for this article were collected through the Multicenter AIDS Cohort Study (MACS), with centers (Principal Investigators) at the Johns Hopkins Bloomberg School of Public Health (Drs. Joseph B. Margolick and Lisa Jacobson), the Howard Brown Health Center and Northwestern University Medical School (Dr. John Phair), the University of California, Los Angeles (Drs. Roger Detels and Beth Jamieson), and the University of Pittsburgh (Dr. Charles Rinaldo), and the Women's Interagency HIV Study (WIHS) Collaborative Study Group, with centers at the New York City/Bronx Consortium (Dr. Kathryn Anastos), Brooklyn, New York (Dr. Howard Minkoff), the Washington, DC, Metropolitan Consortium (Dr. Mary Young), the Connie Wofsy Study Consortium of Northern California (Dr. Ruth Greenblatt), the Los Angeles County/Southern California Consortium (Dr. Alexandra Levine), the Chicago Consortium (Dr. Mardge Cohen), and the Data Coordinating Center (Dr. Stephen Gange). World Wide Web links for both studies are located at http://www.statepi.jhsph.edu.

Conflict of interest: none declared.

REFERENCES

1. Robins JM. Marginal structural models. In: 1997 proceedings of the American Statistical Association, Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association, 1998:1–10. (http://biosun1.harvard.edu/~robins/msm-web.pdf).
2. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health 2006;60:578–86.
3. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60.
4. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000;11:561–70.
5. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of non-randomized treatments. J Am Stat Assoc 2001;96:440–8.
6. Cook NR, Cole SR, Hennekens CH. Use of a marginal structural model to determine the effect of aspirin on cardiovascular mortality in the Physicians' Health Study. Am J Epidemiol 2002;155:1045–53.
7. Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. Stat Med 2002;21:1689–709.
8. Choi HK, Hernán MA, Seeger JD, et al. Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study. Lancet 2002;359:1173–7.
9. Cole SR, Hernán MA, Robins JM, et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. Am J Epidemiol 2003;158:687–94.
10. Ko H, Hogan JW, Mayer KH. Estimating causal treatment effects from longitudinal HIV natural history studies using marginal structural models. Biometrics 2003;59:152–62.
11. Barron Y, Cole SR, Greenblatt RM, et al. Effect of discontinuing antiretroviral therapy on survival of women initiated on highly active antiretroviral therapy. AIDS 2004;18:1579–84.
12. Bodnar LM, Davidian M, Siega-Riz AM, et al. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. Am J Epidemiol 2004;159:926–34.
13. Tager IB, Haight T, Sternfeld B, et al. Effects of physical activity and body composition on functional limitation in the elderly: application of the marginal structural model. Epidemiology 2004;15:479–93.
14. Cole SR, Hernán MA, Margolick JB, et al. Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. Am J Epidemiol 2005;162:471–8.
15. 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;14:477–91.
16. Wang C, Vlahov D, Galai N, et al. The effect of HIV infection on overdose mortality. AIDS 2005;19:935–42.
17. Sterne JA, Hernán MA, Ledergerber B, et al. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. Lancet 2005;366:378–84.
18. Petersen ML, Wang Y, van der Laan MJ, et al. Assessing the effectiveness of antiretroviral adherence interventions. Using marginal structural models to replicate the findings of randomized controlled trials. J Acquir Immune Defic Syndr 2006;43(suppl 1):S96–103.
19. Cole SR, Hernán MA, Anastos K, et al. Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model. Am J Epidemiol 2007;166:219–27.
20. Eisner MD, Wang Y, Haight TJ, et al. Secondhand smoke exposure, pulmonary function, and cardiovascular mortality. Ann Epidemiol 2007;17:364–73.
21. Petersen ML, Wang Y, van der Laan MJ, et al. Pillbox organizers are associated with improved adherence to HIV antiretroviral therapy and viral suppression: a marginal structural model analysis. Clin Infect Dis 2007;45:908–15.
22. Lopez-Gatell H, Cole SR, Hessol NA, et al. Effect of tuberculosis on the survival of women infected with human immunodeficiency virus. Am J Epidemiol 2007;165:1134–42.
23. Cotter D, Zhang Y, Thamer M, et al. The effect of epoetin dose on hematocrit. Kidney Int 2008;73:347–53.
24. Patel K, Hernan MA, Williams PL, et al. Long-term effectiveness of highly active antiretroviral therapy on the survival of children and adolescents with HIV infection: a 10-year follow-up study. Clin Infect Dis 2008;46:507–15.
25. Robins JM. Association, causation, and marginal structural models. Synthese 1999;121:151–79.
26. Brumback BA, Hernán MA, Haneuse SJ, et al. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stat Med 2004;23:749–67.
27. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology 2001;12:313–20.

28. Hernán MA, Hernandez-Diaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155:176–84.
29. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology 2003;14:300–6.
30. Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615–25.
31. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 1986; 123:392–402.
32. Mortimer KM, Neugebauer R, van der Laan M, et al. An application of model-fitting procedures for marginal structural models. Am J Epidemiol 2005;162:382–8.
33. Wang Y, Petersen ML, Bangsberg D, et al. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. Berkeley, CA: Division of Biostatistics, University of California, 2006. (http://www.bepress.com/ucbbiostat/paper211).
34. Brookhart MA, van der Laan M. A semiparametric model selection criterion with applications to the marginal structural model. J Comput Stat Data Anal 2006;50:475–98.
35. Thompson WA Jr. On the treatment of grouped observations in life studies. Biometrics 1977;33:463–70.
36. D'Agostino RB, Lee ML, Belanger AJ, et al. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. Stat Med 1990;9: 1501–15.
37. Steenland K, Deddens JA. A practical guide to dose-response analyses and risk assessment in occupational epidemiology. Epidemiology 2004;15:63–70.
38. Kish L. Weighting for unequal P_i. J Off Stat 1992;8:183–200.
39. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. Comput Methods Programs Biomed 2004;75:45–9.
40. Hernán MA. Invited commentary: hypothetical interventions to define causal effects—afterthought or prerequisite? Am J Epidemiol 2005;162:618–20.; discussion 21–2.
41. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. Int J Obes (in press).
42. Cole SR, Frangakis CE. On the consistency statement in causal inference: a definition or an assumption? Epidemiology (in press).

APPENDIX 1

Inverse Probability Weights

For each subject i, the outcome Y_{ij} was the $\log_{10}$ number of copies of HIV-1 RNA measured in blood at each semiannual study visit j, and the exposure $X_{ij} = 1$ indicates initiation of HAART before visit j, and zero otherwise. We assume that exposure is continuously used after initiation and write $\bar{X}_{ij}$ to denote exposure history up to visit j, that is, $\bar{X}_{ij} = \{X_{i0}, X_{i1}, \ldots, X_{ij}\}$. In our example with 17 follow-up visits beyond baseline, there are 18 possible exposure histories, namely, never initiating HAART, which occurred for 632 (69 percent) of 918 participants, or initiating HAART at any of the 17 follow-up visits. The 286 (31 percent) of 918 participants who initiated HAART did so at visits 1 through 17 in the following numbers: 56, 54, 30, 18, 17, 12, 12, 17, 15, 4, 5, 8, 2, 8, 7, 12, and 9. Here, the structural model is a mapping between each of these 18 static exposure regimes and the mean $\log_{10}$ viral load. The covariate vector $\boldsymbol{L}_{ij}$ available at each visit j included the time-varying covariates CD4 cell count and HIV-1 RNA viral load, as well as the time-fixed (i.e., baseline) covariates sex, race/ethnicity, and age. As with exposure histories, we denote covariate histories by $\bar{\boldsymbol{L}}_{ij}$. Finally, $C_{ij} = 1$ if participant i is censored by visit j, and zero otherwise. Details about the structural (i.e., weighted) model were published previously (19).

The stabilized inverse probability weights SW_{ij} for participant i at visit j are typically the product of inverse probability-of-exposure weights SW^X_{ij} and inverse probability-of-censoring weights SW^C_{ij}; that is, $\mathrm{SW}_{ij} = \mathrm{SW}^X_{ij} \times \mathrm{SW}_{ij}{}^C$. The weight SW^X_{ij} adjusts for measured confounding by the variables in $\bar{\boldsymbol{L}}_{ij}$, and the weight SW^C_{ij} adjusts for measured selection bias by the variables in $\bar{\boldsymbol{L}}_{ij}$. Formally, the component weights are defined as

$$\mathrm{SW}^X_{ij} = \prod_{k=0}^{j} \frac{f[X_{ik}|\bar{X}_{ik-1}, \boldsymbol{V}_{i0}, \bar{C}_{ik-1} = \bar{\mathbf{0}}]}{f[X_{ik}|\bar{X}_{ik-1}, \bar{\boldsymbol{L}}_{ik-1}, \bar{C}_{ik-1} = \bar{\mathbf{0}}]}$$

and

$$\mathrm{SW}^C_{ij} = \prod_{k=1}^{j+1} \frac{\Pr[C_{ik} = 0|\bar{C}_{ik-1} = \bar{\mathbf{0}}, \bar{X}_{ik}, \boldsymbol{V}_{i0}]}{\Pr[C_{ik} = 0|\bar{C}_{ik-1} = \bar{\mathbf{0}}, \bar{X}_{ik}, \bar{\boldsymbol{L}}_{ik-1}]},$$

where $f[\cdot|\cdot]$ is the conditional density function evaluated at the observed covariate values for a given participant, $\bar{\mathbf{0}}$ is a vector of zeros, and $\boldsymbol{V}_{i0}$ is a vector including a subset of the time-fixed baseline variables that is described in more detail below. Note that we ensure the correct temporal order between possible confounders and exposure by using covariate information through visit $j - 1$ $\bar{\boldsymbol{L}}_{ij-1}$, rather than through visit j, to predict exposure reported at visit j, which represents HAART initiation in the interval between visits $j - 1$ and j. Ensuring the proper temporal sequence between confounders and exposure is paramount to the estimation of causal effects, although sometimes published accounts omit any references to this issue.

Bias adjustment is achieved by the denominator of the weights. The numerator of the weights, which does not depend on the time-varying covariates $\bar{\boldsymbol{L}}_{ij}$, is added for stabilization. In many published applications of marginal structural models (9, 14, 19), the conditioning event in the numerator of the weights includes a subset of baseline variables $\boldsymbol{V}_{i0}$ to help stabilize the weights and, thus, obtain narrower confidence intervals around the effect estimate. Informally, to achieve stabilization, one wishes to minimize the difference between the numerator and denominator of the weights such that the remaining difference reflects only the confounding due to the time-varying covariates $\bar{\boldsymbol{L}}_{ij}$, which cannot be appropriately adjusted for by standard regression or stratification. Colloquially, one wishes the numerator of the weight to *chase* the denominator but stop short of following the denominator when it comes to the set of time-varying confounders one wishes to adjust for by weighting.

However, this added stabilization comes at a price: The $\boldsymbol{V}_{i0}$-stabilized weights create a pseudo-population in which,

at each time, exposure is randomized only within the levels of the covariates in $\boldsymbol{V}_{i0}$. In other words, in the pseudo-population, there may still be confounding by $\boldsymbol{V}_{i0}$. Thus, the weighted regression model (equivalently, the marginal structural model) must include the covariates $\boldsymbol{V}_{i0}$ to adjust for this possible confounding. As a result, the estimated causal effect will not be unconditional (marginal) but conditional on the covariates $\boldsymbol{V}_{i0}$. For example, our estimate of a 1.91-$\log_{10}$ decrease in viral load assumes that the effect is the same within levels of baseline variables $\boldsymbol{V}_{i0}$. We could have tested this assumption (of a constant effect across levels of baseline variables $\boldsymbol{V}_{i0}$) by adding product terms to our weighted regression model. Indeed, in several analyses, we have explored possible effect modification by baseline variables (9, 14, 19).

Unstabilized weights, in which the numerator $f[X_{ik}|\bar{X}_{ik-1}, \boldsymbol{V}_{i0}, \bar{C}_{ik-1} = 0]$ is replaced by $f[X_{ik}|\bar{X}_{ik-1}, \bar{C}_{ik-1} = 0]$, can also be used to adjust for bias, but they usually lead to more extreme weights that result in wider confidence intervals around effect estimates. Hence, stabilized weights are generally preferred, even if they require adding the baseline variables in $\boldsymbol{V}_{i0}$ to the weighted model. When one is interested in evaluating potential modification of the exposure effect by the baseline variables, the weighted model must include main effects and product terms for the components of $\boldsymbol{V}_{i0}$, and thus stabilized weights can be used to achieve a greater efficiency at no cost.

APPENDIX 2

Formal Definitions of Identifiability Assumptions

The following three assumptions are needed to nonparametrically identify causal effects. Some methods may not require one or more of these assumptions (e.g., instrumental variables, G-estimation of structural nested models) but, to consistently estimate causal effects, these alternative methods must trade these assumptions for other parametric assumptions.

Consistency means that a subject's counterfactual outcome under her observed exposure history is precisely her observed outcome. To define consistency, let us first define an individual's potential, or counterfactual, outcome $Y_{ij}(\bar{x})$ under exposure history $\bar{x}$ as the outcome that would have been observed if the individual had received exposure history $\bar{x}$. Then, consistency is defined as $Y_{ij}(\bar{x}) = Y_{ij}$ if $\bar{X}_{ij} = \bar{x}$. Our use of the term consistency differs from the statistical property of "consistency," which means that the bias of an estimator approaches zero as information (e.g., sample size) increases. Refer to references 40–42 for a more detailed discussion of consistency for common exposures in epidemiologic research.

Exchangeability states that, given measured confounders $\bar{\boldsymbol{L}}_{ij-1}$, the potential outcomes $Y_{ij}(\bar{x})$ are independent of observed exposure X_{ij} or, in the case of a categorical exposure, $\Pr\left[X_{ij} = x|\bar{\boldsymbol{L}}_{ij-1}, \bar{X}_{ij-1}\right] = \Pr\left[X_{ij} = x|\bar{\boldsymbol{L}}_{ij-1}, \bar{X}_{ij-1}, Y_{ij}(\bar{x})\right]$. In studies with dropout, a similar exchangeability assumption is used for censoring.

Positivity states that there is a nonzero (i.e., positive) probability of receiving every level of exposure X_{ij} for every combination of values of exposure and covariate histories $\bar{X}_{ij-1}$ and $\bar{\boldsymbol{L}}_{ij-1}$ that occur among individuals in the population. Positivity requires that, if $f\left(\bar{X}_{ij-1}, \bar{\boldsymbol{L}}_{ij-1}, \bar{C}_{ij-1} = \bar{\boldsymbol{0}}\right) \neq 0$, then $\Pr\left(X_{ij} = x|\bar{X}_{ij-1}, \bar{\boldsymbol{L}}_{ij-1}, \bar{C}_{ij-1} = \bar{\boldsymbol{0}}\right) > 0$ for all $x \in X_{ij}$. When the analysis uses $\boldsymbol{V}_{i0}$-stabilized weights, as in our case, the positivity condition is slightly weaker because positivity then assumes that, for each value of the baseline covariates $\boldsymbol{V}_{i0}$, there is a nonzero probability of every level of exposure X_{ij} for every combination of values of exposure and covariate histories $\bar{X}_{ij-1}$ and $\bar{\boldsymbol{L}}_{ij-1}$ that occur among individuals with that value of $\boldsymbol{V}_{i0}$. Formally, within levels of $\boldsymbol{V}_{i0}$, positivity requires

$$\frac{\Pr(X_{ij} = x_{ij}|\bar{X}_{ij-1}, \boldsymbol{V}_{i0}, \bar{C}_{ij-1} = \bar{\boldsymbol{0}})}{\Pr(X_{ij} = x_{ij}|\bar{X}_{ij-1}, \bar{\boldsymbol{L}}_{ij-1}, \bar{C}_{ij-1} = \bar{\boldsymbol{0}})} < \infty$$

for all $x \in X_{ij}$, which implies that the assumption holds whenever $\Pr\left(X_{ij} = x_{ij}|\bar{X}_{ij-1}, \boldsymbol{V}_{i0}, \bar{C}_{ij-1} = \bar{\boldsymbol{0}}\right)$ equals zero, regardless of whether $\Pr\left(X_{ij} = x|\bar{X}_{ij-1}, \bar{\boldsymbol{L}}_{ij-1}, \bar{C}_{ij-1} = \bar{\boldsymbol{0}}\right) > 0$.

In fact, our definition of the inverse probability weights in appendix 1 is incomplete: The inverse probability weights are equal to SW_{ij} only under positivity. If the positivity assumption does not hold, then the weights are undefined, and the weights SW_{ij} may result in biased estimates of the causal effect (for details, refer to the Appendix of the paper by Hernán and Robins (2)).