Using Inverse Probability-Weighted Estimators in Comparative Effectiveness Analyses With Observational Databases

Lesley H. Curtis, PhD,† Bradley G. Hammill, MS,* Eric L. Eisenstein, DBA,*† Judith M. Kramer, MD, MS,* and Kevin J. Anstrom, PhD*‡*

Abstract: Inverse probability-weighted estimation is a powerful tool for use with observational data. In this article, we describe how this propensity score-based method can be used to compare the effectiveness of 2 or more treatments. First, we discuss the inherent problems in using observational data to assess comparative effectiveness. Next, we provide a conceptual explanation of inverse probability-weighted estimation and point readers to sources that address the method in more formal, technical terms. Finally, we offer detailed guidance about how to implement the estimators in comparative effectiveness analyses.

Key Words: comparative study, regression analysis, statistical models, propensity score, bias (epidemiology), treatment effects, causal inference, observational studies

(*Med Care* 2007;45: S103–S107)

In clinical and epidemiological studies, researchers often ask whether 1 treatment is better than another at improving survival or preventing disease relapse. For example, in a study of 2 treatments that are randomly assigned—as in a randomized controlled trial—the average causal treatment effect can be defined as the difference between the average response of individuals receiving one of the treatments and the average response of individuals receiving the other treatment ($\bar{X}_1 - \bar{X}_2$). When the treatments are not randomly assigned—as in an observational study—individuals who receive 1 treatment may not be comparable to those receiving the other treatment. To the extent that 1 group is different from the other group in ways that affect the study outcomes (eg, they are sicker, older, poorer, less adherent), any observed difference in outcomes between the 2 groups may simply reflect underlying differences between the groups rather than effects that are caused by the treatments.

Researchers are often most interested in estimating a difference that they cannot observe—what they would have observed had the same individual been exposed to both treatments. The "potential outcomes framework"[1–4] is useful for understanding this unobservable difference. The set of potential outcomes describes the responses and treatments that would have been observed had an individual been subjected to both treatments. For any particular individual, we observe only 1 treatment; the remaining, hypothetical quantity is referred to as the "counterfactual." The average causal treatment effect is defined as:

$$\mu = E\{R^{(1)}\} - E\{R^{(2)}\} \quad (1)$$

where $E\{R^{(1)}\}$ and $E\{R^{(2)}\}$ are the average responses for the entire population if every individual received treatment 1 and treatment 2, respectively.

To make inferences about the distribution of counterfactual responses from the observed data, we must assume that treatment assignment depends only on observed covariates. Stated more formally, this assumption requires that treatment assignment is independent of the counterfactual responses and conditional on observed covariates. In addition, the assumption guarantees that every individual has a positive probability of receiving each treatment.

In comparative effectiveness studies with observational databases, analysts commonly use the so-called propensity score to estimate the average causal treatment effect.[5] The propensity score is the probability of exposure to treatment conditional on observed covariates, and it can be used to balance covariates across treatment groups. Typically, analysts estimate propensity scores from a parametric model such as a logistic regression model, and they compare individuals with similar estimated propensity scores by either stratification or matching. Matching by propensity score controls for many observed covariates simultaneously by matching subjects in 1 treatment group with subjects in another treatment group on the basis of individual propensity scores.[5,6] The difference in average treatment effects between

From the *Duke Clinical Research Institute; and Departments of †Medicine and ‡Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina.

Supported by a Centers for Education and Research on Therapeutics cooperative agreement (3 U18 HS010548-07S1) between Duke University and the Agency for Healthcare Research and Quality.

Reprints: Lesley H. Curtis, PhD, Center for Clinical and Genetic Economics, Duke Clinical Research Institute, PO Box 17969, Durham, NC 27715. E-mail: lesley.curtis@duke.edu.

Presented at the DEcIDE Program Symposium "Comparative Effectiveness and Safety: Emerging Methods" on June 19 and 20, 2006, Gaithersburg, MD.

ISSN: 0025-7079/07/4500-0103

the 2 groups is calculated as the difference in outcomes between the matched groups. With stratification by propensity score, average effect is calculated within each stratum, and the causal difference is estimated as the average of the within-stratum effects. Although the number of strata is left to the discretion of the analyst, stratifying on quintiles is a common practice.[5,7]

Both matching and stratification for the construction of comparison groups have limitations that may constrain their practical application. Matching algorithms frequently omit a significant proportion of the population when comparison groups are being constructed, thus limiting the ability to generalize from the results. Moreover, although stratification will produce treatment groups with similar probabilities for receiving treatment 1 and treatment 2, the individuals in these strata may be indistinguishable to clinicians. Researchers need alternative methods that make more parsimonious use of observational data and that produce analyses that can be applied in clinical decision making.

As an alternative to matching or stratification, Cassel et al[8]; Rosenbaum[9]; and Hirano and Imbens[10] have recommended applying inverse propensity score estimators or inverse probability-weighted estimators to adjust for confounding. Compared with matching and stratification, semiparametric inverse probability-weighted estimators require fewer distributional assumptions about the underlying data, and they avoid the potential residual confounding that arises from stratification on a fixed number of strata.[7] In addition, inverse probability-weighted estimators can incorporate time-dependent covariates and deal with censored data.

In this article, we describe how researchers can use inverse probability-weighted estimators with observational databases to analyze comparative effectiveness. We begin with a conceptual explanation of the estimators and point readers to sources that address the method in more formal, technical terms. Next, we discuss how to use inverse probability-weighted estimators for comparative effectiveness analyses. Finally, we identify priorities and topics for future research.

Inverse Probability-Weighted Estimators

Imagine a sample of data from n patients with treatment indicators (A_i), response variables (R_i), and individual covariates (X_i) assumed to be independent and identically distributed, $i = 1, \ldots, n$. The propensity score typically is unknown and must be estimated based on the observed covariates and treatment assignments. Denote the estimated propensity score as $\pi_a(X_i,\hat{\gamma})$ and $I(\bullet)$ as the treatment indicator function, taking the value 1 if the condition holds and 0 otherwise. The inverse probability-weighted estimate of treatment-specific effect, μ_a, is given by the solution to the following estimating equation:

$$\sum_{i=1}^{n} \frac{I(A_i = a)(R_i - \mu_a)}{\pi_a(X_i,\hat{\gamma})} = 0, a = \{1,2\} \qquad (2)$$

To estimate the causal effect of treatment 1, for example, the analyst includes an observed response (R_i) in the numerator if the individual received treatment 1. (Observed responses for individuals who received treatment 2 are included only in the numerator of the estimate of the causal effect of treatment 2.) The response variable can be generalized in a variety of ways. For example, response may be an estimate of the population mean on the basis of observed covariates or an estimate of the cumulative distribution function (which is useful for estimating cumulative incidence rates in the presence of competing risks).[11] The denominator of the estimating equation is the probability of receiving a given treatment—the propensity score. In cases in which only 2 treatments are possible, only 1 propensity model needs to be fit: an individual's probability of receiving treatment 2 is simply 1 minus the probability of receiving treatment 1. Individuals with a high predicted probability of a given treatment receive a lower weight, compared with individuals with a low predicted probability of a given treatment. Thus, an individual with a low predicted probability of receiving treatment 1, who actually received treatment 1, will represent a larger group of individuals who did not receive treatment 1. (See Robins and Rotnitzky[12] and Robins et al[13] for a detailed exposition of this approach.)

In addition to the relative advantages described earlier, inverse probability-weighted estimators allow researchers to deal with another issue common in observational data—censoring. In observational studies, the response (eg, survival, lifetime medical costs, cumulative hospital admissions) is often observed after some period that may vary by individual. Because of this time lag and the limited follow-up in many studies, some response data will be censored. Although the random timing of an individual's entry into a study (eg, index hospitalization) accounts for much of the variability in time lag, individuals may become censored for other reasons, including changes in health insurance coverage, treatment crossover, and loss to follow-up.

In the equations that follow, we introduce notation for the ascertainment time (T), the potential censoring time (C), and the treatment-specific censoring distribution $\{K_a(t)\}$. Because the true censoring distributions typically are unknown, we must estimate them on the basis of observed data. If we assume that censoring is unrelated to covariates or potential outcomes, we can estimate the censoring distributions using Kaplan–Meier estimates stratified by treatment. Extending the inverse probability-weighted estimate to account for censoring has been described previously in numerous technical publications.[13–16] Briefly, Eq. (2) is expanded to the following:

$$\sum_{i=1}^{n} \frac{I(T_i < C_i)I(A_i = a)(R_i - \mu_a)}{\hat{K}_a(U_i)\pi_a(X_i,\hat{\gamma})} = 0, a = \{1,2\}. \qquad (3)$$

where the treatment-specific censoring distribution $K_a(t)$ typically is estimated using Kaplan–Meier estimates of the censoring distribution. To the numerator the analyst adds an indicator variable to restrict the sample to uncensored individuals—those for whom the study end point was reached before the observation was censored $[I(T_i < C_i)]$. The analyst expands the denominator to include a term reflecting the probability of not being censored.

In more general settings, however, censoring may depend on baseline- and time-dependent variables that, in turn, relate to the individual's response to treatment. A more conservative approach is to assume that the censoring process is conditional on covariate information. Under this assumption, the analyst can estimate the censoring distribution using treatment-specific Cox proportional hazards models.

The response of an uncensored individual to a given treatment, therefore, is inversely weighted by the product of 2 probabilities: the probability of assignment to a given treatment and the probability of being uncensored (eg, the probability of having complete data). Individuals who are less likely to be observed in a given treatment group with complete data (ie, those with low propensity scores who are more likely to be censored) are weighted most heavily.

A drawback of these simple weighted estimators is that only uncensored ("complete") cases are included in the numerator. Suppose we are interested in estimating 3-year medical costs, and an individual is lost to follow-up at 2 years and 11 months. The simple inverse probability-weighted estimators require complete data and would exclude those 2 years and 11 months of cost data. In many studies, including studies involving surgery and expensive one-time therapies, the majority of cost information is obtained in the time period immediately after the initial treatment. It seems reasonable that the information collected from partial observations could be used to construct more efficient estimators. Partitioned estimators, similar to those described by Bang and Tsiatis[16] for randomized studies, allow greater efficiency by incorporating data from these partial observations up to the point of censoring.

The general idea is to divide the follow-up period into nonoverlapping partitions and estimate the average causal treatment effect within each partition. As a result of partitioning the follow-up interval, individuals considered to be censored for the simple weighted estimators may contribute their medical costs for one or more partitions. In general, partitioned estimators will have smaller asymptotic variance than the simple weighted estimators.[13]

The censoring distribution within each partition can be modeled using a Kaplan–Meier estimator, but a more robust partitioned estimator can be constructed with Cox models. As with simple weighted estimators, the Cox version of the partitioned estimator is at least as efficient as the Kaplan–Meier version of the partitioned estimator based on the general theory of inverse probability-weighted estimators.[13]

Inverse Probability-Weighted Estimators in Comparative Effectiveness Analyses

Developing the Propensity Model

The first step in developing the propensity model is to understand and identify clearly the treatment selection process in the context of the clinical questions at hand. First, is it possible that an individual who receives treatment 1 could have received treatment 2? To answer this, we rely on clinical judgment, paying careful attention to contraindications to the treatments of interest. If the potential for receiving either treatment does not exist, then inverse probability-weighted estimation is not appropriate. Second, are there known subgroups that might have different response characteristics? If, for example, individuals with prior exposure to a given treatment are likely to respond differently to the treatments of interest, then the analyst should stratify the sample into clearly defined subgroups before constructing the inverse probability-weighted estimates. If the appropriate subgroups for stratification are not known in advance, the analyst may use Cox proportional hazards models or other regression strategies to identify patient characteristics that are associated with treatment modality. Although these characteristics can then be used to define strata for subsequent analyses, analysts should take care to assure that the resulting strata are clinically meaningful.[17]

Consider, for example, a question concerning the comparative effectiveness of evidence-based β-blockers versus nonevidence-based β-blockers for the treatment of heart failure. Randomized controlled trials have demonstrated that 3 β-blockers improve survival in individuals with heart failure. Although randomized controlled trials have shown several older β-blockers to have a survival benefit for patients with a myocardial infarction, no data exist on whether they confer a survival benefit for patients with heart failure. Physicians are frequently confronted with patients who are taking older β-blockers for the treatment of hypertension and who subsequently develop heart failure. The clinical dilemma is whether all of these patients should be switched to newer (and likely more expensive) β-blockers that have clinical trial evidence of a survival benefit in heart failure, or whether the patients can be maintained on the older β-blockers using the rationale that the beneficial effect of β-blockers in heart failure is a "class effect."

The first question is whether patients who received nonevidence-based β-blockers could have received evidence-based β-blockers. Contraindications are uniform across all β-blockers (evidence-based and nonevidence-based), so this basic assumption is likely met. Next, we consider whether we should define subgroups before constructing inverse probability-weighted estimators. If the outcome of interest is survival after an index hospital admission for heart failure, we may want to stratify the analysis on the basis of β-blocker use before admission, because such prior use is likely to be highly predictive of β-blocker use after discharge. We would then analyze patients taking no β-blockers separately from patients taking β-blockers before the index admission for heart failure. Without stratification by subgroup, the inverse probability-weighted estimators will generate a single answer for the entire population.

The next step is to build a multivariable propensity model that balances the treatment groups with respect to observed baseline confounders. Variable selection for propensity models has received relatively little formal attention, although Rubin and Thomas,[18] Hirano and Imbens,[10] Brookhart et al,[19] and Petersen et al[20] have proposed various strategies. In general, candidate variables for the models should include all baseline covariates that might confound the relationship between treatment and outcome.[18] In the heart failure example, candidate variables might include age, race/ethnicity, and prior comorbidities (eg, ischemic heart disease,

hypertension, cardiovascular disease, chronic obstructive pulmonary disease, diabetes mellitus). In general, parsimony is not a priority unless the exposure of interest is rare and, consequently, the propensity model would have relatively few events per variable.[21] As in any model building exercise, the analyst should explore interactions and the appropriate functional form of continuous variables. In certain situations, limited variable selection may be necessary to stabilize estimates in small data sets.

Assessing the extent to which the model balances the treatment groups is critical. A straightforward approach is to stratify by deciles of predicted probability of treatment (propensity score) and compare baseline characteristics across treatment groups within deciles. Assessing balance is somewhat subjective, but the distribution of baseline covariates between the treatment groups should be similar. If balance is not observed within propensity score strata, then the propensity score model needs to be refined further. In addition, examining the distribution of predicted probabilities (propensity scores) by treatment group is useful. Graphically, the distributions should overlap. Nonoverlapping distributions suggest that one or more baseline covariates are strongly predictive of treatment selection, and analysts should consider performing a stratified analysis in such cases.

As noted in the earlier example, a stratified analysis is appropriate if we expect that defined subgroups will have different response characteristics. In addition, stratification is appropriate if a strong predictor of the propensity to receive treatment exists. Consider the effect of era in the use of drug-eluting stents for the treatment of coronary artery stenosis. Drug-eluting stents prevent restenosis by delivering drugs locally to inhibit scar formation. Randomized controlled trials suggest that, compared with bare metal stents, drug-eluting stents result in 70–80% fewer repeat revascularization procedures. Despite little evidence regarding the long-term effectiveness of drug-eluting stents compared with bare metal stents, drug-eluting stents have become the primary mode of coronary revascularization in the United States. Given the rapid uptake of drug-eluting stents, calendar year would be highly predictive of the propensity to receive a drug-eluting stent. Rather than include year in the model, a better strategy is to define meaningful time periods and estimate comparative effectiveness within those periods. This requires the analyst to make some assumptions about the comparability of individuals over time.

Propensity model development is an iterative process. Researchers should take special care to examine the tails of the distribution of propensity scores and trim extreme weights so that they do not exert undue influence. Researchers should also use sensitivity analyses to explore the robustness of the model to the inclusion or exclusion of covariates and alternate specifications.

Modeling the Censoring Distribution

The first step in modeling the censoring distribution is to understand the source of the censoring. Is it administrative in nature, reflecting random entry times and a calendar-based study end, or is the censoring related to baseline- and time-dependent covariates that may be related to an individual's response to treatment? If the censoring is unrelated to an individual's covariates, treatment-specific Kaplan–Meier estimates of the censoring distributions will suffice. If, by contrast, censoring is related to an individual's covariates, a Cox proportional hazards model should be used. Again, modeling the censoring distribution with a Cox model serves primarily to reduce bias. In theory, even if the Kaplan–Meier estimate is sufficient, a Cox model would improve statistical efficiency (eg, smaller variances).[13] Based on our experience, however, unless the Cox model is extended to include post-baseline information, the gains in statistical efficiency are minimal.

The next step in the modeling exercise is to assess the extent of censoring. As noted previously, a partitioned estimator will be more efficient in most cases. The key decision with a partitioned estimator is where to set the partitions. In general, setting the partitions where the analyst expects to collect the most information is best. If the outcomes of interest occur at the beginning of follow-up, for example, setting partitions early in the follow-up phase to capture important data is the best strategy.

Consider, for example, the source and extent of censoring in an analysis of Medicare claims data. Index events occur at random times throughout the study period so that individuals with index events in later years have shorter follow-up. This censoring is administrative in nature and is unrelated to individual covariates. Another form of censoring arises when Medicare beneficiaries switch their coverage from fee-for-service to Medicare managed care. The data are censored during periods of managed care eligibility but, unlike the previous example, the censoring is related to individual covariates (eg, younger age, fewer chronic diseases, and lower expenses in the year before enrollment) (S.V. Rao, et al, unpublished data, 2005). In such situations, a partitioned inverse probability-weighted estimator—in which the censoring distribution is estimated using a Cox model that includes baseline covariates—may be most appropriate. Again, the placement and length of the partitions will depend on the specific clinical question and when the outcomes of interest are most likely to occur.

Other issues arise when modeling the censoring distribution. First, individuals can be lost to follow-up, at least temporarily. This happens, for example, when interviewers are unable to locate individuals for a regularly scheduled follow-up, but reestablish contact at a later date. Similarly, individuals may switch in and out of health plans and, as a result, be missing in a given health plan's data set during the switches out. (This is particularly an issue with Medicare beneficiaries who are allowed to switch between fee-for-service and Medicare managed care on a monthly basis.) Although one can account for this noncontinuous follow-up, in most cases it is reasonable to consider individuals censored when the first censoring event occurs.

In addition, observational data sets are compiled from multiple sources, including billing data, medical claims, and direct individual follow-up. Data from electronic sources (eg, billing data, electronic medical records) can be added to the data set on a weekly (or even daily) basis, whereas individ-

ual-reported data may be collected only once or twice a year. Therefore, hospital-based outcomes are likely to be reflected in the data set soon after they occur. Outcomes or events that occur outside of the hospital (death or rehospitalization at another hospital) will not be reflected in the data set until after the individual reports them. If data are collected from individuals (eg, at 6 and 12 months), then data in a given interval can be considered complete only after the individual-level data collection for that period is finished.[22]

Modeling the censoring distribution is an iterative process. As with propensity model development, it is important to check the distribution of the weights that result from the model. A good general practice with partitioned models is to sum the weights at the end of each partition and verify that the sum of the weights is approximately equal to the initial sample size. If that is not the case, then the analyst should consider trimming extreme weights. Ultimately, the analyst must verify that the final results (ie, the estimated differences in response) are robust to different specifications of these weights.

Calculating Standard Errors

In addition to using analytical formulas, researchers can use 2 methods to calculate standard errors of inverse probability-weighted estimates. The first is to use the robust standard errors generated by the weighted analytic model, but further empirical work is needed to better understand the behavior of these standard error estimates. The second option is to estimate standard errors using the nonparametric bootstrap method,[23] as suggested by Hernan et al.[24] Although intuitively appealing, the bootstrap method is computationally intensive even for moderately sized data sets because, in practice, hundreds of iterations are advisable.

Priorities and Topics for Further Research

Inverse probability-weighted estimation is a powerful tool for use with observational data. Currently a major drawback is the lack of statistical software for implementing these methods. Therefore, the behavior of these estimators will need to be evaluated carefully on a case-by-case basis. Given the flexibility of these methods, however, we expect to see increasing use of inverse probability-weighted estimators in the medical and epidemiology literature.

ACKNOWLEDGMENTS

The authors thank Damon Seils for assistance with manuscript preparation.

REFERENCES

1. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81: 945–960.
2. Splawa-Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat Sci.* 1990;8:465–480. Originally published, in Polish: *Roczniki Nauk Rolniczych*. 1923; 10:1–51.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688–701.
4. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat*. 1978;6:34–58.
5. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a nonrandomized control group. *Stat Med*. 1998;17:2265–2281.
6. Rosenbaum PR, Rubin DB. Constructing a control-group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39:33–38.
7. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23:2937–2960.
8. Cassel CM, Sarndal CE, Wretman JH. Some uses of statistical models in connection with the nonresponse problem. In: Madow WG, Olkin I, eds. *Incomplete Data in Sample Surveys, Vol 3: Symposium on Incomplete Data, Proceedings*. New York, NY: Academic Press; 1983.
9. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82:387–394.
10. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcomes Res Methodol*. 2001;2:259–278.
11. Gooley TA, Leisenring W, Crowley J, et al. Estimation of cumulative incidence in the presence of competing risks. In: Crowley D, Ankerst DP, eds. *Handbook of Statistics in Clinical Oncology.* 2nd ed. Boca Raton, FL: Chapman & Hall; 2005:557–565.
12. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell N, Dietz K, Farewell V, eds. *AIDS Epidemiology: Methodological Issues*. Boston, MA: Birkhauser; 1992:24–33.
13. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846–866.
14. Anstrom KJ, Tsiatis AA. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics*. 2001;57: 1207–1218.
15. Zhao SZ, Tsiatis AA. Efficient estimation of the distribution of quality adjusted survival time. *Biometrics*. 1999;55:1101–1107.
16. Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika*. 2000;87:329–343.
17. Vonesh EF, Snyder JJ, Foley RN, et al. The differential impact of risk factors on mortality in hemodialysis and peritoneal dialysis. *Kidney Int*. 2004;66:2389–2401.
18. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics*. 1996;52:249–264.
19. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149–1156.
20. 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:S96–S103.
21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–45.
22. Fine JP, Tsiatis AA. Testing for differences in survival with delayed ascertainment. *Biometrics*. 2000;56:145–153.
23. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* New York, NY: Chapman & Hall; 1993.
24. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11:561–570.

