BIOMETRICS 57, 1207–1218
December 2001

Utilizing Propensity Scores to Estimate Causal Treatment Effects with Censored Time-Lagged Data

Kevin J. Anstrom* and Anastasios A. Tsiatis

Department of Statistics, North Carolina State University,
Raleigh, North Carolina 27695, U.S.A.
**email:* kjanstro@stat.ncsu.edu

SUMMARY. Observational studies frequently are conducted to compare long-term effects of treatments. Without randomization, patients receiving one treatment are not guaranteed to be prognostically comparable to those receiving another treatment. Furthermore, the response of interest may be right-censored because of incomplete follow-up. Statistical methods that do not account for censoring and confounding may lead to biased estimates. This article presents a method for estimating treatment effects in nonrandomized studies with right-censored responses. We review the assumptions required to estimate average causal effects and derive an estimator for comparing two treatments by applying inverse weights to the complete cases. The weights are determined according to the estimated probability of receiving treatment conditional on covariates and the estimated treatment-specific censoring distribution. By utilizing martingale representations, the estimator is shown to be asymptotically normal and an estimator for the asymptotic variance is derived. Simulation results are presented to evaluate the properties of the estimator. These methods are applied to an observational data set of acute coronary syndrome patients from Duke University Medical Center to estimate the effect of a treatment strategy on the mean 5-year medical cost.

KEY WORDS: Censoring; Confounding; Inverse weighting; Kaplan–Meier; Observational study; Propensity score; Survival analysis.

1. Introduction

In many biomedical studies, the primary aim is to estimate the difference in mean response between two treatments. However, the primary response variable, R, often is not available immediately after the patient enters the study. Rather, the response is observed after some period of time that may vary by individual. The length of time before the response is observed, which is denoted by the positive random variable T, is called the lag-time or time to response ascertainment and we will refer to R as the time-lagged response. Examples of time-lagged responses include survival time, lifetime medical costs, quality-adjusted survival time, and cumulative hospital admissions. Because of the time lag and the fact that many studies have limited follow-up, some of the response data will be censored.

In observational studies, because patients are not randomized to treatment, the data analyst must be careful in dealing with potential confounding. The propensity score, which is the probability that an individual is assigned one of the treatments as a function of observed covariates, is commonly employed to adjust for confounding in large nonrandomized studies (Rubin, 1997; Dehejia and Wahba, 1999). Typically, the propensity scores are estimated from a parametric model and individuals with similar estimated propensity scores are compared either by stratification or matching. Alternatively, Cassel, Särndal, and Wretman (1983) and Rosenbaum (1987) suggested inverse propensity score estimators to adjust for confounding. The topic of this paper will be to generalize the inverse propensity score estimators to the setting with censored data.

In the next section, we review the assumptions required to estimate average causal effects and describe the inverse propensity score estimators when there is no censoring. We extend the problem to deal with right censoring of time-lagged responses in Section 3. The large sample properties of the estimator are outlined in Section 4. In Section 5, we conduct numerical studies to evaluate the small sample properties of the proposed estimator under various censoring and confounding patterns. In Section 6, we estimate treatment effects from an observational study of patients with coronary artery disease.

2. Estimating Causal Effects with Propensity Scores

We begin by reviewing the definition of average causal treatment effect when there is no censoring. With no censoring, the response R will be known for all the individuals in the study. Thus, the lag-time variable T need not be considered for the time being. Throughout the paper we consider counterfactual random variables (or potential outcomes) as described by Rubin (1974, 1978). We define $R^{(0)}$ to correspond to the response of a randomly chosen individual in our population if, possibly contrary to fact, the patient received treatment 0. Similarly, we define $R^{(1)}$ as the response if the patient re-

ceived treatment 1. These are hypothetical quantities because an individual can receive only one treatment. Nonetheless, the average causal treatment effect is defined as

$$\delta = E(R^{(1)}) - E(R^{(0)}).$$

In actuality, the experimental sample will receive (be assigned) only one treatment, 0 or 1, and this will be denoted by the treatment indicator $A = (0, 1)$ and the observed response $R = R^{(0)}I(A = 0) + R^{(1)}I(A = 1)$. It is important to understand what conditions are required to identify the average causal treatment effect from the distribution of the observable random variables (R, A). For example, in a randomized study, it is reasonable to assume that $(R^{(0)}, R^{(1)}) \perp\!\!\!\perp A$, where $\perp\!\!\!\perp$ denotes being statistically independent. Under this assumption, the average causal treatment effect can be expressed in terms of the population parameters for the observable random variables (R, A), namely,

$$\delta = E(R \mid A = 1) - E(R \mid A = 0).$$

In an observational study, patients receiving treatment 1 may not be comparable to those receiving treatment 0; therefore, it may not be reasonable to assume that $(R^{(0)}, R^{(1)}) \perp\!\!\!\perp A$. However, if prognostic factors X can be identified that are believed to explain the prognostic variation and if, in addition, we believe that treatment choice only depends on X, then conditional on X, it may be reasonable to assume that treatment assignment is random. We denote the independence of the counterfactuals and treatment assignment conditional on the vector of observed covariates X by:

$$\left(R^{(0)}, R^{(1)}\right) \perp\!\!\!\perp A \mid X. \tag{1}$$

With this assumption, the average causal treatment effect δ can be identified in terms of the distribution of the observable random variables (R, A, X). This follows because

$$\begin{aligned} E(R^{(1)}) &= E\{E(R^{(1)} \mid X)\} \\ &= E\{E(R^{(1)} \mid A = 1, X)\} \\ &= E\{E(R \mid A = 1, X)\}. \end{aligned}$$

The second equality follows from (1). A similar argument gives $E(R^{(0)}) = E\{E(R \mid A = 0, X)\}$.

In observational studies with large numbers of covariates, adjustment for baseline differences using covariance methods might be inadequate. An alternative strategy to adjust for covariate imbalance is through the use of the propensity score. Proposed by Rosenbaum and Rubin (1983), the propensity score is defined as the probability of being assigned treatment 1 (say) conditional on X:

$$\pi(x) = P(A = 1 \mid X = x).$$

Propensity scores are particularly useful because they allow the data analyst to adjust for prognostic differences while reducing the dimension of the covariates.

We say that treatment assignment is strongly ignorable if we observe covariates X, such that $(R^{(0)}, R^{(1)}) \perp\!\!\!\perp A \mid X$ and $0 < \pi(X) < 1$ (Rosenbaum and Rubin, 1983; Rosenbaum, 1984). In addition to the independence between the counterfactual responses and treatment assignment conditional on covariates, the strong ignorability assumption guarantees that every individual has a positive probability of receiving either treatment. If we believed the strong ignorability assumption and either knew or could estimate the propensity score, then we could take advantage of the relationship

$$E\left\{\frac{I(A = 1)(R - \mu_1)}{\pi(X)}\right\} = 0, \tag{2}$$

where $\mu_1 = E(R^{(1)})$, to obtain estimators.

Equation (2) follows because

$$\begin{aligned} &E\left\{\frac{I(A = 1)(R - \mu_1)}{\pi(X)}\right\} \\ &\quad = E\left\{\frac{I(A = 1)(R^{(1)} - \mu_1)}{\pi(X)}\right\} \\ &\quad = E\left[E\left\{\frac{I(A = 1)(R^{(1)} - \mu_1)}{\pi(X)} \mid R^{(1)}, X\right\}\right] \\ &\quad = E\left[\frac{(R^{(1)} - \mu_1)}{\pi(X)} E\{I(A = 1) \mid R^{(1)}, X\}\right] \\ &\quad = E(R^{(1)} - \mu_1) = 0. \end{aligned}$$

The next to last step follows from the strong ignorability assumption. Similarly,

$$E\left[\frac{I(A = 0)(R - \mu_0)}{\{1 - \pi(X)\}}\right] = 0$$

where $\mu_0 = E(R^{(0)})$.

Suppose we had a sample of data $(R_i, A_i, X_i), i = 1, \ldots, n$ assumed independent and identically distributed, then an estimator for μ_1 is given by the solution to the estimating equation $\Sigma_{i=1}^{n} A_i(R_i - \tilde{\mu}_1)/\pi(X_i) = 0$ or equivalently

$$\tilde{\mu}_1 = \frac{\displaystyle\sum_{i=1}^{n} \frac{A_i R_i}{\pi(X_i)}}{\displaystyle\sum_{i=1}^{n} \frac{A_i}{\pi(X_i)}}. \tag{3}$$

Similarly an estimator for μ_0 is given by

$$\tilde{\mu}_0 = \frac{\displaystyle\sum_{i=1}^{n} \frac{(1 - A_i)R_i}{1 - \pi(X_i)}}{\displaystyle\sum_{i=1}^{n} \frac{1 - A_i}{1 - \pi(X_i)}}. \tag{4}$$

For observational studies the propensity score is generally unknown and must be estimated from the observed data (A_i, X_i). For this purpose, we might assume a parametric model such as $P(A = 1 \mid X = x) = \pi(x, \gamma)$, where γ denotes a finite set of unknown parameters. Because treatment is binary, the logistic regression model is often used for this purpose. That is

$$\log\left[\frac{\pi(x, \gamma)}{\{1 - \pi(x, \gamma)\}}\right] = x^T\gamma.$$

We do not need to restrict ourselves to only these models. However, we will consider estimating the parameter γ

using maximum likelihood, which we denote by $\hat{\gamma}$, and we will assume the necessary regularity conditions so that this estimator will be asymptotically normal. We denote $\pi(x, \gamma_t)$ to be the true propensity score that generated the data and $\pi(x, \hat{\gamma})$ to be the maximum likelihood estimate of the propensity score, where $\pi(x, \gamma) = \exp(x^T\gamma)/[1 + \exp(x^T\gamma)]$. By replacing $\pi(X_i)$ with the estimate $\pi(X_i, \hat{\gamma})$ in (3) and (4), an estimator for average causal treatment effect is given by

$$\tilde{\delta} = \frac{\sum_{i=1}^{n} \frac{A_i R_i}{\pi(X_i, \hat{\gamma})}}{\sum_{i=1}^{n} \frac{A_i}{\pi(X_i, \hat{\gamma})}} - \frac{\sum_{i=1}^{n} \frac{(1 - A_i)R_i}{1 - \pi(X_i, \hat{\gamma})}}{\sum_{i=1}^{n} \frac{1 - A_i}{1 - \pi(X_i, \hat{\gamma})}}. \tag{5}$$

These estimators were originally suggested for missing data problems (Cassel et al., 1983; Rosenbaum, 1987) and are similar to the Horvitz–Thompson estimator (Horvitz and Thompson, 1952) in the survey sampling context. Little (1986) discussed other weighting strategies based on the stratification of estimated propensity scores.

3. Estimating Causal Effects in Right Censored Data with Propensity Scores

As stated earlier, in many biomedical applications the response variable of interest may be censored. Let C_i denote the ith individual's potential censoring time, in which case the observable data are defined by

U_i = $\min(T_i, C_i)$ = time to response ascertainment or censoring,

Δ_i = $I\{(T_i \le C_i)\}$ = complete-case indicator,

and

R_i = response observed only if $\Delta_i = 1$.

Now that censoring has been introduced, we need to define counterfactual lag-times $\{T_i^{(0)}, T_i^{(1)}\}$ that correspond to the lag-times if, contrary to fact, the ith individual in our sample received treatment 0 or 1, respectively. The observed lag-time $T_i = T_i^{(0)} I(A_i = 0) + T_i^{(1)} I(A_i = 1)$. The assumption of strong ignorability is now extended to

$$\{T_i^{(0)}, T_i^{(1)}, R_i^{(0)}, R_i^{(1)}\} \perp\!\!\!\perp A_i \mid X_i$$

and

$$0 < \pi(X_i) < 1.$$

In addition, we assume that censoring is noninformative conditional on treatment assignment; specifically,

$$C_i \perp\!\!\!\perp \{T_i^{(0)}, T_i^{(1)}, R_i^{(0)}, R_i^{(1)}, X_i\} \mid A_i \tag{6}$$

which also implies that $C_i \perp\!\!\!\perp (T_i, R_i, X_i) \mid A_i$. This assumption may be weakened to allow

$$C_i \perp\!\!\!\perp \{T_i^{(0)}, T_i^{(1)}, R_i^{(0)}, R_i^{(1)}\} \mid (A_i, X_i), \tag{7}$$

but for many problems the first assumption will often suffice. Let us denote the treatment-specific censoring distributions as $K_1(u)$ and $K_0(u)$, where

$$K_j(u) = P(C \ge u \mid A = j), \qquad j = 0, 1.$$

We must also assume that the lag-time T is bounded, say, by the value L, and that $K_1(L)$ and $K_0(L)$ are bounded away from zero. This assumption is necessary to guarantee that there is some probability of observing individuals with all possible lag-times.

The observable data are denoted by the sample of i.i.d. random vectors

$$(U_i, \Delta_i, \Delta_i R_i, A_i, X_i), \qquad i = 1, \ldots, n.$$

The focus is to use this data to estimate $E(R^{(0)})$, $E(R^{(1)})$, and the average causal treatment effect $\delta = E(R^{(1)}) - E(R^{(0)})$. The proposed estimator will be an inverse-probability–weighted estimator motivated by the following relationship:

$$E\left\{\frac{I(A = 1)\Delta(R - \mu_1)}{\pi(X)K_1(U)}\right\} = E(R^{(1)} - \mu_1) = 0. \tag{8}$$

We obtain this relationship by noting

$$\begin{aligned} &E\left\{\frac{I(A = 1)\Delta(R^{(1)} - \mu_1)}{\pi(X)K_1(T^{(1)})}\right\} \\ &\quad = E\left[E\left\{\frac{I(A = 1)I(C \ge T^{(1)})(R^{(1)} - \mu_1)}{\pi(X)K_1(T^{(1)})} \mid \right.\right. \\ &\qquad\qquad \left.\left. T^{(1)}, R^{(1)}, X, A\right\}\right] \\ &\quad = E\left[\frac{I(A = 1)(R^{(1)} - \mu_1)}{\pi(X)K_1(T^{(1)})}\right. \\ &\qquad\qquad \left. \times E\left\{I(C \ge T^{(1)}) \mid T^{(1)}, R^{(1)}, X, A\right\}\right]. \end{aligned}$$

The inner expectation, given as $K_1(T^{(1)})I(A = 1) + K_0(T^{(1)}) \times I(A = 0)$, when multiplied by $I(A = 1)$ yields $K_1(T^{(1)}) \times I(A = 1)$. Therefore, the left hand side of (8) equals

$$\begin{aligned} &E\left\{\frac{I(A = 1)(R^{(1)} - \mu_1)K_1(T^{(1)})}{\pi(X)K_1(T^{(1)})}\right\} \\ &\qquad = E\left\{\frac{I(A = 1)(R^{(1)} - \mu_1)}{\pi(X)}\right\}, \end{aligned}$$

which by equation (2) equals 0. Similarly,

$$E\left[\frac{I(A = 0)\Delta(R - \mu_0)}{\{1 - \pi(X)\}K_0(U)}\right] = E(R^{(0)} - \mu_0) = 0.$$

In general, $\pi(x)$, $K_0(u)$, and $K_1(u)$ are not known and must be estimated. In Section 2, we briefly discussed estimating the propensity score using maximum likelihood. We also propose estimating $K_1(u)$ and $K_0(u)$ with treatment-specific Kaplan–Meier estimators (Kaplan and Meier, 1958) of the censoring distribution. These estimators would be obtained by stratifying on treatment and then reversing the role of failure and censoring to obtain the Kaplan–Meier estimator for the censoring distribution. We denote the treatment-specific censoring estimators as $\hat{K_1}(u)$ and $\hat{K_0}(u)$. The estimator we propose for δ is given as

$$\hat{\delta} = \hat{\mu_1} - \hat{\mu_0}, \tag{9}$$

where $\hat{\mu_1}$ and $\hat{\mu_0}$ are solutions to

$$\sum_{i=1}^{n} \frac{A_i\Delta_i(R_i - \hat{\mu_1})}{\pi(X_i,\hat{\gamma})\hat{K}_1(U_i)} = 0$$

and

$$\sum_{i=1}^{n} \frac{(1-A_i)\Delta_i(R_i - \hat{\mu_0})}{\{1-\pi(X_i,\hat{\gamma})\}\hat{K}_0(U_i)} = 0,$$

respectively. Similar approaches using inverse weights have been suggested to estimate survival distributions (Hubbard, van der Laan, and Robins, 1999) and parameters in marginal structural models (Hernàn, Brumback, and Robins, 2001).

The estimator (9) provides the data analyst with a tool to potentially estimate causal effects from data sets with censored data. We would like to note that this estimator is not efficient among the class of semiparametric estimators that inversely weight complete cases and assume that censoring follows relationship (6). Further work needs to be done to obtain more efficient semiparametric estimators, including methods for using partial response information from individuals who are censored to gain efficiency. However, the proposed estimator provides easily computed consistent estimates and avoids the difficult task of correctly specifying the direct relationship between the time-lagged response and censoring given the covariates. According to Rubin (1997), an advantage of propensity score methods compared to standard statistical models is that they allow the data analyst to examine the overlap of the covariate distributions. Without adequate covariate overlap, estimating causal effects must rely on model-dependent extrapolations.

4. Large Sample Properties

The strategy taken for showing asymptotic normality is to derive the influence function of the proposed estimator. That is, we show that the estimator minus the estimand can be approximated by a sum of independent mean zero random variables. Namely,

$$n^{1/2}(\hat{\delta} - \delta) = n^{-1/2}\sum_{i=1}^{n}\psi_i + o_p(1),$$

where ψ_i is a function of the ith individual's data, such that $E(\psi_i) = 0$ and $E(\psi_i^2)$ is bounded and $o_p(1)$ corresponds to a term that converges in probability to zero as n goes to infinity. The random variable ψ_i is defined as the ith influence function of the estimator. If we can identify the influence function of $\hat{\delta}$, then we immediately know that the estimator is asymptotically normal. In addition, the asymptotic variance of $\hat{\delta}$ is given as $E(\psi^2)$.

Clearly, the influence function of $\hat{\delta}$ is the difference of the influence functions of $\hat{\mu_1}$ and $\hat{\mu_0}$. If the propensity score $\pi(x)$ and the censoring distribution $K_1(u)$ are known, then the influence function for $\hat{\mu_1}$ can be easily obtained by noting that

$$\begin{aligned} & n^{1/2}(\hat{\mu_1} - \mu_1) \\ &= \frac{n^{-1/2}\sum_{i=1}^{n}\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i)K_1(U_i)}}{n^{-1}\sum_{i=1}^{n}\frac{A_i\Delta_i}{\pi(X_i)K_1(U_i)}} \\ &= n^{-1/2}\sum_{i=1}^{n}\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i)K_1(U_i)} + o_p(1). \end{aligned}$$

The last equality above follows from the fact that

$$n^{-1}\sum_{i=1}^{n}\frac{A_i\Delta_i}{\pi(X_i)K_1(U_i)}$$

converges in probability to $E\{A_i\Delta_i/\pi(X_i)K_1(U_i)\}$, which can be shown to equal one using a conditioning argument similar to (8). Consequently, if the propensity score and censoring distribution were known, the ith influence function for $\hat{\mu_1}$ would equal $A_i\Delta_i(R_i - \mu_1)/\pi(X_i)K_1(U_i)$. Similarly, the ith influence function for $\hat{\mu_0}$ would equal $(1-A_i)\Delta_i(R_i-\mu_0)/\{1-\pi(X_i)\}K_0(U_i)$.

The technical difficulty comes from the fact that $\hat{\delta}$ involves estimated propensity scores, $\pi(x,\hat{\gamma})$, and estimated censoring distributions, $\hat{K_0}(u)$ and $\hat{K_1}(u)$. Therefore, in order to derive the influence function, we must account for the influence of estimating the parameter γ in the propensity score by the maximum likelihood estimator $\hat{\gamma}$ and estimating the treatment-specific censoring distribution by the corresponding Kaplan–Meier estimator.

In the Appendix, we give the details in calculating the ith influence function of $\hat{\delta}$ and show that $n^{1/2}(\hat{\delta}-\delta)$ is asymptotically normal. The asymptotic variance of $n^{1/2}(\hat{\delta}-\delta)$ can be consistently estimated by

$$\begin{aligned} \hat{\sigma}^2 = & \sum_{i=1}^{n}\frac{\Delta_i}{n\hat{K}_{A_i}(U_i)} \\ & \times\left[\frac{A_i(R_i-\hat{\mu_1})}{\pi(X_i,\hat{\gamma})} - \frac{(1-A_i)(R_i-\hat{\mu_0})}{\{1-\pi(X_i,\hat{\gamma})\}}\right. \\ & \left. - \hat{H}^T\{\hat{E}(S_\gamma S_\gamma^T)\}^{-1}X_i\{A_i-\pi(X_i,\hat{\gamma})\}\right]^2 \\ & + \sum_{i=1}^{n}\frac{A_i(1-\Delta_i)\{\hat{G}_1(U_i)\}}{\left\{\sum_{l=1}^{n}A_lY_l(U_i)\right\}\hat{K}_1(U_i)} \\ & + \sum_{i=1}^{n}\frac{(1-A_i)(1-\Delta_i)\{\hat{G}_0(U_i)\}}{\left\{\sum_{l=1}^{n}(1-A_l)Y_l(U_i)\right\}\hat{K}_0(U_i)}, \end{aligned} \tag{10}$$

where

$$Y_i(x) = I(U_i \geq x),$$

$$\hat{G}_0(u) = \sum_{i=1}^{n}\frac{(1-A_i)\Delta_i}{n\hat{K}_0(U_i)}$$

$$\times\left[\frac{(R_i-\hat{\mu_0})}{\{1-\pi(X_i,\hat{\gamma})\}}\right.$$

$$- \frac{\sum_{j=1}^{n} \dfrac{(1-A_j)\Delta_j(R_j-\hat{\mu}_0)I(U_j \geq u)}{\{1-\pi(X_j,\hat{\gamma})\}\hat{K}_0(U_j)}}{\sum_{j=1}^{n} \dfrac{(1-A_j)\Delta_j I(U_j \geq u)}{\hat{K}_0(U_j)}} \Bigg]^2 \times I(U_i \geq u),$$

$$\hat{G}_1(u) = \sum_{i=1}^{n} \frac{A_i\Delta_i}{n\hat{K}_1(U_i)} \times \Bigg[\frac{(R_i-\hat{\mu}_1)}{\pi(X_i,\hat{\gamma})} - \frac{\sum_{j=1}^{n} \dfrac{A_j\Delta_j(R_j-\hat{\mu}_1)I(U_j \geq u)}{\pi(X_j,\hat{\gamma})\hat{K}_1(U_j)}}{\sum_{j=1}^{n} \dfrac{A_j\Delta_j I(U_j \geq u)}{\hat{K}_1(U_j)}} \Bigg]^2 \times I(U_i \geq u),$$

$$\hat{H} = \frac{\sum_{i=1}^{n} \left[\dfrac{A_i\Delta_i\{R_i-\hat{\mu}_1\}X_i\{1-\pi(X_i,\hat{\gamma})\}}{\hat{K}_1(U_i)\pi(X_i,\hat{\gamma})} \right]}{\sum_{i=1}^{n} \dfrac{A_i\Delta_i}{\hat{K}_1(U_i)\pi(X_i,\hat{\gamma})}} + \frac{\sum_{i=1}^{n} \left[\dfrac{(1-A_i)\Delta_i\{R_i-\hat{\mu}_0\}X_i\pi(X_i,\hat{\gamma})}{\hat{K}_0(U_i)\{1-\pi(X_i,\hat{\gamma})\}} \right]}{\sum_{i=1}^{n} \dfrac{(1-A_i)\Delta_i}{\hat{K}_0(U_i)\{1-\pi(X_i,\hat{\gamma})\}}},$$

and

$$\hat{E}(S_\gamma S_\gamma^T) = n^{-1}\sum_{i=1}^{n} X_i X_i^T \pi(X_i,\hat{\gamma})\{1-\pi(X_i,\hat{\gamma})\}.$$

5. Simulations

We conducted simulation studies to evaluate the properties of the proposed estimator, $\hat{\delta}$. Specifically, we considered a situation where the primary goal was to estimate the difference in mean medical costs between two treatments in an observational study when some of the data are censored. Our simulations reflect the notion of counterfactual responses, as described previously. Namely, if individuals were assigned treatment 0, then medical costs, denoted by $R^{(0)}$, would be incurred over a period $T^{(0)}$ (lag-time). Similarly, if individuals were assigned to treatment 1, then medical costs $R^{(1)}$, would be incurred over a period $T^{(1)}$ In our simulations, imbalance of prognostic factors was reflected by introducing two covariates, denoted by X_1 and X_2, that affect treatment assignment, medical costs, and time lags.

We considered three sets of simulation scenarios. For the first two simulation scenarios, the two covariates X_1 and X_2 were generated as independent standard normal random variables truncated at ± 1.96 and the counterfactual lag-time variables were generated as

$$T^{(j)} = \min\{4, \exp(0.5X_1 + 0.5X_2 + \epsilon)\}, \qquad j = 0, 1, \quad (11)$$

where ϵ was a standard normal random variable. Thus, the lag-time was bounded by a maximum value of $L = 4$. Additionally, we considered the counterfactual medical costs for treatment 0 to be

$$R^{(0)} = e^{8+Z_1} + e^{6+Z_2}T^{(0)}\max(X_1, 0) + e^{5+Z_3}T^{(0)}\max(X_2, 0), \quad (12)$$

where Z_1, Z_2, and Z_3 were independent standard normal random variables. The random variable $R^{(0)}$ represents a situation where there was an initial cost at time 0 plus additional costs, depending on the lag-time and the values of the covariates. For the first two simulations, the treatment assignment $A = (0, 1)$ for each individual was generated as independent Bernoulli random variables with $\Pr(A = 1 \mid X_1, X_2) = (e^{\gamma_0+\gamma_1X_1+\gamma_2X_2})/(1+e^{\gamma_0+\gamma_1X_1+\gamma_2X_2})$. Different values of γ_1 and γ_2 were chosen to study varying amounts of confounding.

The total medical cost of patients may not always be observed because of incomplete follow-up or censoring. Specifically, the time lag may be censored by the variable C, which was generated according to four censoring patterns. For the first two patterns, the censoring distributions were identical for both treatments. The uniform $(0, 6)$ and uniform $(2, 6)$ censoring distributions resulted in approximately 25% and 10% of the simulated data being censored, respectively. For the last two patterns, treatment 0 had uniform $(0, 6)$ and uniform $(2, 6)$ censoring, whereas treatment 1 had no censoring.

In the simulation experiments, not only did we want to study the properties of the proposed estimator, but we also wanted to consider the consequences of analyses that didn't properly account for either confounding or censoring. Therefore, we considered three estimators. The proposed estimator $\hat{\delta}$ was calculated using (9), where $\hat{K}_1(u)$ and $\hat{K}_0(u)$ were treatment-specific Kaplan–Meier estimates of the censoring survival distributions and $\hat{\pi}(x)$ was estimated with a logistic regression model that included an intercept term along with X_1 and X_2. Next, we wanted to consider an estimator that accounted for censoring but not for confounding. A one-sample estimator for censored mean medical costs was given by Bang and Tsiatis (2000). Treatment difference, which does not account for covariate imbalance, was estimated simply as the difference in these treatment-specific estimators. This estimator, denoted by $\hat{\delta}_{\text{int}}$, was identical to using estimator (9) and estimating the propensity score with a logistic regression model with only an intercept term. Finally, a naive approach that did not account for censoring was to consider only complete cases and delete all observations that were censored. This estimator, denoted by $\hat{\delta}_{\text{cc}}$, was calculated by applying (5) to the complete cases ($\Delta = 1$) and estimating $\hat{\pi}(x)$ with a logistic regression model using an intercept term, X_1 and X_2. We expected $\hat{\delta}_{\text{cc}}$ to adjust for covariate imbalances but not for censoring.

Each simulation scenario used 2000 replications. The results were summarized by the empirical bias (BIAS) and empirical standard error (Sim SE) of the estimates across the replications. The accuracy of our estimate for the standard error was evaluated by comparing the empirical standard error to the empirical average of the estimated standard errors

Table 1
Simulation summary for the medical cost data with $R^{(1)} = R^{(0)}$

	$\gamma_1 = \gamma_2 =$	0.00		0.25		0.50		0.75		1.00	
	Sample Size:	500	1000	500	1000	500	1000	500	1000	500	1000
Censoring pattern 1: uniform (2, 6) for both treatments											
$\hat{\delta}$	BIAS	26.3	1.6	−0.3	−7.3	9.5	7.0	34.2	3.4	−12.8	−2.6
$\hat{\delta}_{cc}$	BIAS	24.3	2.0	−4.1	−9.6	11.1	7.6	24.0	0.1	−19.1	−1.9
$\hat{\delta}_{int}$	BIAS	28.0	1.3	263	265	520	513	710	685	790	818
$\hat{\delta}$	Sim SE	665	460	677	477	715	495	783	556	940	635
$\hat{\delta}$	Ave SE	641	460	645	468	679	488	721	532	786	590
$\hat{\delta}$	ECP	.949	.954	.950	.946	.947	.956	.940	.956	.936	.945
Censoring pattern 2: uniform (0, 6) for both treatments											
$\hat{\delta}$	BIAS	19.3	1.7	−18.2	−21.9	16.5	9.2	23.8	−3.6	−20.1	−7.1
$\hat{\delta}_{cc}$	BIAS	21.6	1.5	−13.5	−21.7	17.5	5.6	19.2	−7.3	−30.0	−9.8
$\hat{\delta}_{int}$	BIAS	22.3	1.6	248	253	530	513	704	680	789	818
$\hat{\delta}$	Sim SE	779	539	801	547	839	569	912	656	1103	722
$\hat{\delta}$	Ave SE	722	520	728	530	761	551	806	599	873	657
$\hat{\delta}$	ECP	.940	.948	.940	.950	.938	.956	.934	.945	.933	.943
Censoring pattern 3: uniform (2, 6) for treatment 0, none for treatment 1											
$\hat{\delta}$	BIAS	22.4	3.6	1.6	−5.4	8.7	6.3	26.7	3.1	−7.7	0.2
$\hat{\delta}_{cc}$	BIAS	131	113	108	101	123	121	136	115	99	112
$\hat{\delta}_{int}$	BIAS	24.1	3.1	265	268	520	512	702	686	798	823
$\hat{\delta}$	Sim SE	634	439	658	454	693	479	764	540	923	621
$\hat{\delta}$	Ave SE	616	441	624	452	658	474	702	518	770	579
$\hat{\delta}$	ECP	.945	.957	.952	.949	.948	.955	.941	.950	.941	.947
Censoring pattern 4: uniform (0, 6) for treatment 0, none for treatment 1											
$\hat{\delta}$	BIAS	14.7	4.2	−17.5	−13.1	0.7	9.0	15.8	−2.0	−21.0	0.8
$\hat{\delta}_{cc}$	BIAS	169	155	142	139	168	169	184	159	138	157
$\hat{\delta}_{int}$	BIAS	16.9	3.6	250	263	512	513	695	683	790	824
$\hat{\delta}$	Sim SE	703	480	736	506	776	528	857	609	1047	689
$\hat{\delta}$	Ave SE	659	473	672	489	705	510	753	561	822	622
$\hat{\delta}$	ECP	.936	.951	.939	.945	.939	.947	.934	.944	.926	.939

For all studies, the mean of $R^{(0)}$ was approximately 5650 with a CV of 1.2.

(Ave SEs) across the replications. Finally, the large-sample accuracy was evaluated by examining the empirical coverage probability (ECP) of the 95% confidence intervals; that is, the proportion of the replications where the estimate was within 1.96 estimated standard errors from the truth.

For the first simulation experiment, we examined the effect of the treatment assignment parameters $(\gamma_0, \gamma_1, \gamma_2)$, sample size, and the censoring distribution on the estimators. We took $\gamma_0 = 0$ and let γ_1 and γ_2 vary, taking on the values of 0.00, 0.25, 0.50, 0.75, and 1.00. In each simulated data set, approximately 50% of individuals received treatment 1. We considered samples of size 500 and 1000. We set each individual's counterfactual response for treatment 1 to be identical to their counterfactual response for treatment 0. Namely, for the ith individual, $R_i^{(1)} = R_i^{(0)}$, where $R^{(0)}$ is defined in (12). Table 1 contains a summary of the simulation results. As expected, the proposed estimator, $\hat{\delta}$, with the correct model for the propensity score, was nearly unbiased and the 95% coverage probabilities had nearly their nominal level for all parameter settings. Typically, the Sim SE was larger than the Ave SE. However, with samples of size 1000, the two measures of standard error never differed by more than 10%. The complete-case estimator, $\hat{\delta}_{cc}$, was nearly unbiased for the parameter δ when the censoring distributions were identical for both treatments. However, $\hat{\delta}_{cc}$ was biased for censoring patterns 3 and 4. As expected $\hat{\delta}_{int}$, the intercept-only estimator, which did not account for confounding, was biased except when $\gamma_1 = \gamma_2 = 0$.

The second simulation study considered the scenario where $R^{(1)}$ differed from $R^{(0)}$ by an additive factor multiplied by the lag-time. Specifically, $R^{(1)} = R^{(0)} + \alpha T^{(1)}$, where $\alpha = 200, 400, 600, 800, 1000$. In Table 2, we see that the proposed estimator $\hat{\delta}$ was nearly unbiased with 95% coverage probabilities ranging from 93% to 95%. The Ave SE was on average about 5% smaller than the simulation standard error. As α increased, the bias of $\hat{\delta}_{cc}$ increased. At $\alpha = 1000$, the bias of $\hat{\delta}_{cc}$ was −380 compared to −18 for the proposed estimator. The intercept-only estimator, $\hat{\delta}_{int}$, was very biased, with 95% coverage probabilities ranging from 73% to 82%. The true value of δ was too difficult to evaluate analytically and there-

Table 2
Simulation summary with additive treatment effects[a]

$R^{(1)} = R^{(0)} + \alpha T^{(1)}$				
Estimators		$\hat{\delta}$	$\hat{\delta}_{cc}$	$\hat{\delta}_{int}$
$\alpha = 200$	BIAS	−10	−79	534
$\delta = 298$	Sim SE	581	513	537
	Ave SE	557	503	525
	ECP	.952	.951	.816
$\alpha = 400$	BIAS	10	−140	593
$\delta = 596$	Sim SE	558	495	525
	Ave SE	550	503	523
	ECP	.951	.949	.797
$\alpha = 600$	BIAS	13	−218	643
$\delta = 894$	Sim SE	598	512	548
	Ave SE	550	501	524
	ECP	.944	.927	.754
$\alpha = 800$	BIAS	−4	−313	666
$\delta = 1192$	Sim SE	630	551	574
	Ave SE	561	509	530
	ECP	.930	.882	.735
$\alpha = 1000$	BIAS	−18	−380	700
$\delta = 1490$	Sim SE	594	523	552
	Ave SE	562	509	532
	ECP	.945	.879	.730

[a] Two thousand Monte Carlo simulations were used for each study. Censoring was generated as uniform (0, 6) for both treatments. Sample sizes were 1000. Treatment assignment parameters were $\gamma_0 = 0$, $\gamma_1 = \gamma_2 = 1/2$. The true causal treatment effect was denoted by δ. For all studies the mean of $R^{(0)}$ was approximately 5650 with a CV of 1.2.

fore was approximated by averaging 500,000 replicates. We also considered the case where $R^{(1)}$ differed from $R^{(0)}$ by a multiplicative factor and observed similar qualitative results.

A third simulation study, suggested by a referee, was designed to examine the effect of unequal treatment-specific covariate variances on the proposed estimator. The treatment assignment for each individual was generated as independent Bernoulli random variables with $\Pr(A = 1) = 0.50$. For individuals with treatment assignment $A = 0$, the covariates X_1 and X_2 were generated as independent normal random variables with mean γ and variance 2, where $\gamma = 0$ and 0.25. For individuals with treatment assignment $A = 1$, the covariates X_1 and X_2 were generated as independent normal random variables with mean $-\gamma$ and variance 3. The counterfactual lag-times and medical costs were generated according to (11) and (12), respectively. For this setup, the true propensity score model that is induced is a polynomial logistic model that includes quadratic terms and interactions of the covariates (Rosenbaum and Rubin, 1983). The propensity score models used for estimators $\hat{\delta}$ and $\hat{\delta}_{cc}$ were fitted using two different logistic regression models. In the first case, a logit model with only linear terms was used. The second model was a polynomial logit model including linear, quadratic, and interaction terms. We would expect the polynomial logit model to yield less biased results because this model represents the correct specification of the propensity score. We also examined the bias of these estimators for estimating the parameter μ_1. The results of the simulation are shown in Table 3. The proposed estimator with the polynomial logit model $\hat{\delta}^{p}$ was nearly unbiased for both δ and μ_1 and had estimated 95% coverage between 93% and 94%. The large difference between the simulation standard error and the average standard error was primarily due to a few outliers. For the linear logit model, the estimator $\hat{\delta}^{l}$ was more biased and had 95% coverage probabilities of approximately 90%. Both complete case estimators, $\hat{\delta}^{l}_{cc}$ and $\hat{\delta}^{p}_{cc}$, were biased for parameter μ_1 in all settings. Additionally, $\hat{\delta}^{l}_{cc}$ was biased for δ. As expected, the intercept-only estimator, $\hat{\delta}_{int}$, was biased and had poor coverage probabilities for both parameter settings.

Table 3
Simulation summary with unequal covariate variances[a]

$R^{(1)} = R^{(0)}$						
Estimators		$\hat{\delta}^{l}$	$\hat{\delta}^{p}$	$\hat{\delta}^{l}_{cc}$	$\hat{\delta}^{p}_{cc}$	$\hat{\delta}_{int}$
$\gamma = 0$	BIAS(δ)	−374	6	−285	4	−376
	BIAS(μ_1)	−184	−1	−700	−576	−185
	Sim SE	629	686	515	534	641
	Ave SE	611	629	509	512	622
	ECP	.901	.939	.922	.945	.902
$\gamma = 0.25$	BIAS(δ)	−395	−31	−284	10	−1144
	BIAS(μ_1)	−289	−27	−794	−615	−577
	Sim SE	642	823	521	607	648
	Ave SE	616	691	512	541	632
	ECP	.882	.936	.913	.948	.542

[a] The propensity score models for $\hat{\delta}^{l}$ and $\hat{\delta}^{l}_{cc}$ included an intercept term, X_1, and X_2. The propensity score models for $\hat{\delta}^{p}$ and $\hat{\delta}^{p}_{cc}$ included an intercept term, X_1, X_2, an interaction term, and second-order terms. Two thousand Monte Carlo simulations were used for each study. Censoring was uniform (0, 6) for both treatments. Sample sizes were 1000 with $\Pr(A = 1) = 0.50$. The covariates X_1 and X_2 were independent normal random variables. For $A = 0$, the mean of the covariates was γ and the variance was 2. For $A = 1$, the mean of the covariates was $-\gamma$ and the variance was 3. The true causal treatment effect was zero. For both studies, the mean of $R^{(0)}$ was approximately 6400 with a CV of 1.2.

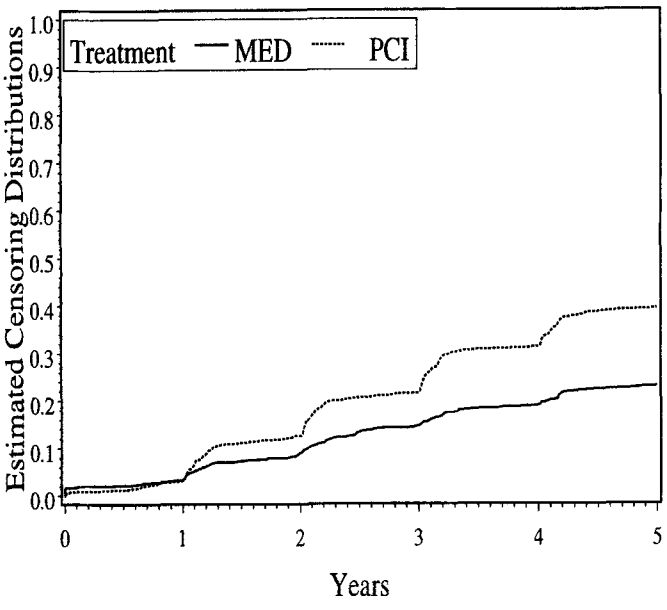

Figure 1. Estimated treatment-specific censoring distributions where the curves represent one minus the Kaplan–Meier estimates for censoring.

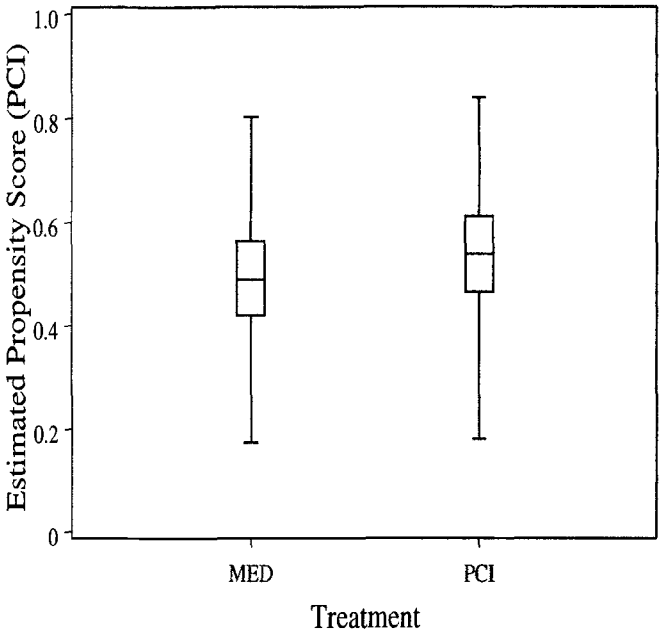

Figure 2. Estimated treatment-specific distributions of propensity scores (for PCI treatment) with the final model.

6. Example

As part of an observational study to estimate the economic burden of acute coronary syndromes on a population of patients with heart disease, we applied the proposed method to estimate 5-year medical cost treatment differences. A Duke University Medical Center database prospectively recorded baseline demographics, medical history, and catheterization results for patients with acute coronary syndrome who received their initial heart catheterization between 1986 and 1997. The purpose of the analysis was to describe medical cost differences between an initial treatment of coronary angioplasty (PCI) and medicine (MED) (see Mark et al., 1994). The study population included individuals with one- or two-vessel coronary artery disease, ejection fraction $\geq$ 30%, and no history of congestive heart failure. Attempts were made to contact patients 6 months after their initial heart catheterization and annually thereafter. A total of 1657 patients received an initial treatment strategy of PCI and 1557 patients received MED. All costs were converted to 1997 dollar values and discounted at a rate of 3% per year (Eisenstein et al., 2001).

For this analysis, an individual was considered to have complete cost data (i.e., $\Delta_i = 1$) if they were followed for 5+ years or if they died before the end of 1998. A total of 2284 out of 3214 patients have complete data. Among the individuals with complete cost data, the mean medical cost for an initial treatment of PCI was \$41,793 compared to \$26,801 for MED, for a difference of \$14,992. As one might have expected, the medical cost data were right-skewed. For PCI patients with complete cost data, the median cost was \$32,226, whereas the 95th and 99th percentiles of cost were \$104,105 and \$170,971. For MED patients with complete cost data, the median cost was \$16,219, whereas the 95th and 99th percentiles of cost were \$80,442 and \$173,204.

To apply the proposed estimator, we must estimate the censoring distributions and propensity scores. The treatment-specific censoring distributions were estimated with Kaplan–Meier estimates and are shown in Figure 1. The probability of being censored by five years was 0.39 for the PCI patients compared to 0.23 for the MED patients. The difference in the censoring distributions was at least partially due to the increased use of PCI during the study period (Mark et al., 1994).

As a preliminary analysis, a propensity score model was constructed using a logistic regression model. Twelve variables were considered for the propensity score model. These variables were age (median 59, range 30–91), sex, race, ejection fraction (median 55, range 30–92), hypertension, history of myocardial infarction, mitral insufficiency, diabetes, peripheral vascular disease, unstable angina status, history of smoking, and number of diseased coronary vessels (one or two). We also considered all two-way interactions, and higher-order terms for the continuous variables. An initial logit model was constructed using a stepwise variable selection technique. Additional terms were added to the model following the iterative procedure described by Rosenbaum and Rubin (1984). The final propensity score model contained 23 terms. The three strongest predictors in the propensity score model based on the chi-squared statistic were hypertension, history of smoking, and ejection fraction. Individuals with higher ejection fraction measurements were more likely to receive PCI treatment, whereas individuals with a history of smoking or hypertension were more likely to receive MED. Figure 2 displays the distributions of the estimated propensity scores by treatment. The estimated propensity score distributions were quite similar for both treatments. Of the 51% of individuals that actu-

ally received PCI, the estimated propensity scores ranged from a minimum of .18 to a maximum of .84, with a median of .54. By comparison, among those patients receiving MED, the minimum, maximum, and median estimated propensity scores (for PCI treatment) were .17, .80, and .49.

Applying these estimated quantities to (9), the estimated causal treatment effect was \$15,395, with a standard error of \$1314. The estimated 5-year medical cost for an initial treatment of PCI was \$41,736 compared to \$26,341 for MED. Similar treatment effects were obtained by $\hat{\delta}_{cc}$ and $\hat{\delta}_{int}$. The complete-case estimator $\hat{\delta}_{cc}$, estimated the treatment effect to be \$15,797, whereas the $\hat{\delta}_{int}$ estimated the treatment effect to be \$14,591.

We conducted several sensitivity analyses on this data. To assess the impact of model selection, we used a logistic regression model with the 12 variables entered as linear effects. The estimated treatment difference was \$15,446, only \$51 greater than the estimated difference given by the primary model used. Also, to examine the effect of extremely high-cost individuals, we considered log costs. We estimated, with methods described above, the average log cost for PCI and MED to be 10.45 and 9.74, respectively. When exponentiated, these quantities yielded estimates of \$34,201 and \$16,970, for a difference of \$17,231. These figures were similar to the exponentiated (average log cost) values among the complete cases of \$34,372 for PCI and \$17,223 for MED, for a difference of \$17,149.

In this example, the adjustments for confounding and censoring only altered the results slightly. Several reasons are possible for this finding. For this study population, the majority of costs were incurred during the first few months after catheterization, which would dampen the effect of ignoring censoring. Moreover, the factors most predictive in the propensity score model were not very prognostic of medical costs. Nonetheless, for observational data of this type, we believe it is important to conduct such an analysis to ensure that the biases that may have resulted were taken into account.

7. Discussion

In this paper we present a method to estimate causal treatment effects in observational studies using propensity scores. We also consider some of the large sample properties of the estimator and give an estimate for the asymptotic standard error. Extensive simulation studies show that our estimator performs well with moderate sample sizes similar to those used in clinical trial settings. In the previous sections, we have focused on estimating medical cost differences between two treatment strategies. More generally, these methods may be used to estimate treatment effects for any time-lagged response. However, one cautionary note is that inverse-probability–weighted estimators are known to be unstable when the weights are large. In this paper, the weights are a function of the estimated propensity scores and estimated censoring distributions. To bound the estimated censoring distributions away from zero, we suggested using an artificial maximum lag-time. In applications, particular care needs to be taken to restrict attention to those individuals whose covariate values indicate that they have a reasonable chance to receive either treatment. When applying weighting methods to response data with extremely long tails, we suggest that the data analyst conduct some exploratory analyses to examine the sensitivity of the results to a small number of extreme outliers.

Acknowledgements

This research was supported by grant AI-31789 from the National Institute of Allergy and Infectious Disease and grant CA-51962 from the National Cancer Institute. We are very grateful to Eric Eisenstein and the Duke Clinical Research Institute for permission to use the Burden of Illness data set.

Résumé

Les études observationnelles sont fréquemment menées pour comparer les effets au long terme de traitements. Hors randomisation, les patients recevant un des traitements ne présentent plus de comparabilité pronostique par rapport aux patients recevant un autre traitement. De plus la réponse étudiée peut être censurée à droite en raison de suivis interrompus. Les méthodes statistiques ne prenant pas en compte la censure et la non séparation d'effets peuvent ainsi conduire à des estimations biaisées. Cet article présente une méthode pour estimer les effets de traitements dans des études non randomisées, avec réponses censurées à droite. Nous réexaminons les hypothèses nécessaires à l'estimation des effets directs moyens, et nous construisons un estimateur pour la comparaison de deux traitements en appliquant une pondération inverses aux cas complets. Les poids sont déterminés en fonction de la probabilité inverse de recevoir le traitement, conditionnellement aux covariables et à la distribution estimée de la censure spécifique au traitement. En utilisant les martingales, nous montrons que l'estimateur est asymptotiquement normal et nous obtenons une estimation de la variance asymptotique. On présente des résultats de simulation évaluant les propriétés de l'estimateur. Les méthodes sont appliquées à des données observationnelles de patients avec insuffisance coronaire aiguë, du Centre médical de Duke University, pour estimer l'effet d'une stratégie de traitement sur le coût médical moyen à 5 ans.

References

Bang, H. and Tsiatis, A. A. (2000). Estimating medical costs with censored data. *Biometrika* **87,** 329–343.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys*, Volume III, *Symposium on Incomplete Data, Proceedings* W. G. Madow and I. Olkin (eds), 143–160. New York: Academic.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94,** 1053–1062.

Eisenstein, E. E., Shaw, L. K., Anstrom, K. J., Nelson, C. L., Hakim, Z., Hasselblad, V., and Mark, D. B. (2001). Assessing the clinical and economic burden of coronary artery disease: 1986–1998. *Medical Care* **39,** 824–835.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Gill, R. D. (1980). *Censoring and Stochastic Integrals,* Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.

Hernàn, M., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* **96,** 440–448.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663–685.

Hubbard, A. E., van der Laan, M. J., and Robins, J. M. (1999). Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and covariates in observational studies. In *Observational Studies, Statistical Models in Epidemiology, the Environment and Clinical Trials (IMA Volumes in Mathematics and Its Applications)*, M. E. Halloran and D. Berry (eds), 135–178. New York: Springer-Verlag.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* **54,** 139–157.

Mark, D. B., Nelson, C. L., Califf, et al. (1994). Continuing evolution of therapy for coronary-artery disease—initial results from the era of coronary angioplasty. *Circulation* **89,** 2015–2025.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology—Methodological Issues*, N. Jewell, K. Dietz, and V. Farewell (eds), 297–331. Boston: Birkhäuser.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association* **79,** 41–48.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82,** 387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification of the propensity score. *Journal of the American Statistical Association* **79,** 516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **6,** 34–58.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127,** 757–763.

Received November 2000. *Revised April* 2001.
Accepted July 2001.

Appendix 1

Much of the theoretical development relies on martingale methods applied to censoring problems as described in Fleming and Harrington (1991). Let the filtration $\mathcal{F}(u)$ be the increasing sequence of σ-algebras generated by

$$\sigma\{I(C_i \le t),\ t \le u, A_i, X_i, R_i, T_i,\ i = 1, \ldots, n\},$$

which is the σ-algebra generated by all covariates, treatment assignments, response variables, and lag-times, along with censoring history up to time u. We construct an $\mathcal{F}(x)$ martingale process for individual j by defining

$$M_j^c(x) = N_j^c(x) - \int_0^x \left\{\lambda_0^c(t)I(A_j = 0) + \lambda_1^c(t)I(A_j = 1)\right\} Y_j(t)dt,$$

where $N_j^c(x) = I(U_j \le x, \Delta_j = 0)$ is the counting process that counts whether individual j was censored before or at time x, $Y_j(x) = I(U_j \ge x)$ is the indicator of whether individual j is at risk at time x, and $\lambda_0^c(x), \lambda_1^c(x)$ are the censoring hazard functions for treatment groups 0 and 1, respectively. Using results of Gill (1980), the Kaplan–Meier estimator for censoring from treatment group 1 can be represented as

$$\left\{\frac{\hat{K}_1(u) - K_1(u)}{K_1(u)}\right\} = -\sum_{j=1}^{n}\left\{\int_0^u \frac{A_j dM_j^c(x)}{\sum_{l=1}^n A_l Y_l(x)}\right\} + o_p(1). \tag{13}$$

The estimator $\hat{\gamma}$ maximizes the objective function given by

$$L(\gamma, X, A) = \prod_{i=1}^{n}\{\pi(X_i, \gamma)\}^{A_i}\{1 - \pi(X_i, \gamma)\}^{(1-A_i)}.$$

The score vector S_γ, defined as

$$\left.\frac{\partial \log L(\gamma, X, A)}{\partial \gamma}\right|_{\gamma=\gamma_t},$$

is given by $\sum_{i=1}^n S_\gamma(X_i, A_i, \gamma_t)$, where

$$S_\gamma(X_i, A_i, \gamma_t) = \frac{\dfrac{\partial \pi(X_i, \gamma_t)}{\partial \gamma}\{A_i - \pi(X_i)\}}{\pi(X_i)\{1 - \pi(X_i)\}}$$

and γ_t is the true parameter generating the data. If γ is suppressed in $\pi(X_i, \gamma)$, then it is implicitly assumed that $\pi(X_i) = \pi(X_i, \gamma_t)$. Standard asymptotic theory allows us to represent the estimator $\hat{\gamma}$ as follows:

$$n^{1/2}(\hat{\gamma} - \gamma_t) = n^{-1/2}\sum_{i=1}^{n}\{E(S_\gamma S_\gamma^T)\}^{-1} S_\gamma(X_i, A_i, \gamma_t) + o_p(1), \tag{14}$$

where $S_\gamma(X_i, A_i, \gamma_t)$ are independent and identically distributed random variables. By the central limit theorem $\hat{\gamma}$ is unconditionally asymptotically normal with mean zero and covariance matrix $\{E(S_\gamma S_\gamma^T)\}^{-1}$.

We assume that there exists $\epsilon > 0$ such that $K_1(L) > \epsilon, K_0(L) > \epsilon$, and $\epsilon < \pi(X) < (1 - \epsilon)$ with probability 1. The estimator minus the estimand can be represented as $\sqrt{n}(\hat{\delta} - \delta) = \sqrt{n}(\hat{\mu}_1 - \mu_1) - \sqrt{n}(\hat{\mu}_0 - \mu_0)$. We note that

$$\sqrt{n}(\hat{\mu}_1 - \mu_1) = \frac{n^{-1/2}\displaystyle\sum_{i=1}^{n}\left\{\frac{A_i\Delta_i(R_i - \mu_1)}{\pi(X_i, \hat{\gamma})\hat{K}_1(U_i)}\right\}}{n^{-1}\displaystyle\sum_{i=1}^{n}\left\{\frac{A_i\Delta_i}{\pi(X_i, \hat{\gamma})\hat{K}_1(U_i)}\right\}}$$

and under the assumed regularity conditions,

$$n^{-1}\sum_{i=1}^{n}\left\{\frac{A_i\Delta_i}{\pi(X_i,\hat{\gamma})\hat{K}_1(U_i)}\right\}$$

converges in probability to

$$E\left\{\frac{A_i\Delta_i}{\pi(X_i)K_1(U_i)}\right\}=1.$$

Therefore, $\sqrt{n}(\hat{\mu}_1-\mu_1)$ is asymptotically equivalent to

$$n^{-1/2}\sum_{i=1}^{n}\left\{\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\hat{\gamma})\hat{K}_1(U_i)}\right\}. \tag{15}$$

By adding and subtracting common terms (15) can be represented as

$$n^{-1/2}\sum_{i=1}^{n}\left\{\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\hat{\gamma})\hat{K}_1(U_i)}-\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\gamma_t)\hat{K}_1(U_i)}\right\} \tag{16}$$

$$+n^{-1/2}\sum_{i=1}^{n}\left\{\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\gamma_t)\hat{K}_1(U_i)}-\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\gamma_t)K_1(U_i)}\right\} \tag{17}$$

$$+n^{-1/2}\sum_{i=1}^{n}\left\{\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\gamma_t)K_1(U_i)}\right\}. \tag{18}$$

By applying representation of $\hat{\gamma}$ shown in (14), we can expand (16) as

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\left\{\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\hat{\gamma})\hat{K}_1(U_i)}-\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\gamma_t)\hat{K}_1(U_i)}\right\}\\
&=\left\{-n^{-1}\sum_{i=1}^{n}\frac{A_i\Delta_i(R_i-\mu_1)\dfrac{\partial\pi(X_i)^T}{\partial\gamma}}{K_1(U_i)\pi(X_i)\pi(X_i)}\right\}\\
&\quad\times n^{1/2}(\hat{\gamma}-\gamma_t)+o_p(1)\\
&=n^{-1/2}\sum_{i=1}^{n}\left[-E\left\{\frac{(R^{(1)}-\mu_1)\dfrac{\partial\pi(X_i)^T}{\partial\gamma}}{\pi(X_i)}\right\}\right.\\
&\qquad\times\{E(S_\gamma S_\gamma^T)\}^{-1}\\
&\qquad\left.\times\frac{\dfrac{\partial\pi(X_i)}{\partial\gamma}\{A_i-\pi(X_i)\}}{\pi(X_i)\{1-\pi(X_i)\}}\right]+o_p(1).
\end{aligned}$$

Using Gill's representation (13) of the Kaplan-Meier estimator, (17) is equal to

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i,\gamma_t)}\left\{\frac{1}{\hat{K}_1(U_i)}-\frac{1}{K_1(U_i)}\right\}\\
&=-n^{-1/2}\sum_{i=1}^{n}\frac{A_i\Delta_i(R_i-\mu_1)}{\pi(X_i)K_1(U_i)}\left\{\frac{\hat{K}_1(U_i)-K_1(U_i)}{K_1(U_i)}\right\}\\
&\quad+o_p(1)\\
&=n^{-1/2}\sum_{j=1}^{n}\int_0^L\frac{A_j dM_j^c(u)}{\left\{\sum_{l=1}^{n}A_lY_l(x)/n\right\}}\\
&\qquad\times n^{-1}\left\{\sum_{i=1}^{n}\frac{A_i\Delta_i(R_i^{(1)}-\mu_1)}{\pi(X_i)K_1(T_i^{(1)})}I(T_i^{(1)}\geq u)\right\}\\
&\quad+o_p(1)\\
&=n^{-1/2}\sum_{i=1}^{n}\int_0^L\frac{A_i dM_i^c(u)}{K_1(u)}\\
&\qquad\times\left[\frac{E\left\{(R^{(1)}-\mu_1)I(T^{(1)}\geq u)\right\}}{E\left\{AI(T^{(1)}\geq u)\right\}}\right]\\
&\quad+o_p(1).
\end{aligned}$$

Finally, we consider the sum of independent and identically distributed terms (18). Here, we utilize an equality from Robins and Rotnitzky (1992),

$$\frac{\Delta_i}{K(U_i)}=1-\int_0^\infty\frac{dM_i^c(u)}{K(u)}.$$

Thus, the sum of i.i.d. terms (18) can be written as

$$\begin{aligned}
&\sum_{i=1}^{n}\frac{\Delta_iA_i(R_i-\mu_1)}{\pi(X_i)K_1(U_i)}\\
&=\sum_{i=1}^{n}\left[\left\{\frac{A_i(R_i^{(1)}-\mu_1)}{\pi(X_i)}\right\}\right.\\
&\qquad\left.-\int_0^L\frac{A_i(R_i^{(1)}-\mu_1)}{\pi(X_i)}\frac{dM_i^c(u)}{K_1(u)}\right].
\end{aligned}$$

Consequently, the influence function for $\hat{\mu}_1$ can be expressed as

$$\begin{aligned}
&\left\{\frac{A_i(R_i^{(1)}-\mu_1)}{\pi(X_i)}\right\}-E\left\{\frac{(R^{(1)}-\mu_1)\dfrac{\partial\pi(X_i)}{\partial\gamma}}{\pi(X_i)}\right\}^T\\
&\quad\times\left\{E(S_\gamma S_\gamma^T)\right\}^{-1}\frac{\dfrac{\partial\pi(X_i)}{\partial\gamma}\{A_i-\pi(X_i)\}}{\pi(X_i)\{1-\pi(X_i)\}}\\
&\quad-\int_0^L\frac{A_i dM_i^c(u)}{K_1(u)}\\
&\qquad\times\left[\frac{\left\{R_i^{(1)}-\mu_1\right\}}{\pi(X_i)}-\frac{E\left\{(R^{(1)}-\mu_1)I(T^{(1)}\geq u)\right\}}{E\left[AI\left\{T^{(1)}\geq u\right\}\right]}\right].
\end{aligned}$$

By analogy, we can derive the influence function for $\hat{\mu}_0$. Thus, the influence function for $\hat{\delta}=\hat{\mu_1}-\hat{\mu_0}$ is given by

$$\begin{aligned}
\psi_i=&\left\{\frac{A_i(R_i^{(1)}-\mu_1)}{\pi(X_i)}\right\}-\left[\frac{(1-A_i)(R_i^{(0)}-\mu_0)}{\{1-\pi(X_i)\}}\right]\\
&-E\left[\frac{(R_i^{(1)}-\mu_1)\frac{\partial\pi(X_i)}{\partial\gamma}}{\pi(X_i)}+\frac{(R_i^{(0)}-\mu_0)\dfrac{\partial\pi(X_i)}{\partial\gamma}}{\{1-\pi(X_i)\}}\right]^T
\end{aligned}$$

$$\times \left\{E(S_\gamma S_\gamma^T)\right\}^{-1} \frac{\frac{\partial \pi(X_i)}{\partial \gamma}\{A_i - \pi(X_i)\}}{\pi(X_i)\{1-\pi(X_i)\}} \tag{19}$$

$$- \int_0^L \frac{A_i dM_i^c(u)}{K_1(u)} \times \left[\frac{\left\{R_i^{(1)} - \mu_1\right\}}{\pi(X_i)} - \frac{E\left\{(R^{(1)} - \mu_1) I(T^{(1)} \geq u)\right\}}{E\left\{AI(T^{(1)} \geq u)\right\}}\right] \tag{20}$$

$$+ \int_0^L \frac{(1-A_i) dM_i^c(u)}{K_0(u)} \times \left[\frac{(R_i^{(0)} - \mu_0)}{\{1-\pi(X_i)\}} - \frac{E\left\{(R^{(0)} - \mu_0) I(T^{(0)} \geq u)\right\}}{E\left\{(1-A) I(T^{(0)} \geq u)\right\}}\right]. \tag{21}$$

Because we have identified the influence function, we know that the estimator is asymptotically normal with asymptotic variance of $E(\psi^2)$. Because (19) is $\mathcal{F}(0)$ measurable, this means that (19) is uncorrelated with (20) and (21). The covariance between (20) and (21) must be zero because $A_i(1 - A_i) = 0$. Consequently, the variance of the influence function is given as the variance(19) + variance(20) + variance(21).

A consistent estimator of the variance of (19) is given by

$$n^{-1} \sum_{i=1}^n \frac{\Delta_i}{\hat{K}_{A_i}(U_i)} \left[\frac{A_i(R_i - \hat{\mu}_1)}{\pi(X_i, \hat{\gamma})} - \frac{\{1-A_i\}(R_i - \hat{\mu}_0)}{\{1-\pi(X_i, \hat{\gamma})\}} - \hat{H}^T \left\{\hat{E}(S_\gamma S_\gamma^T)\right\}^{-1} X_i \times \{A_i - \pi(X_i, \hat{\gamma})\}\right]^2,$$

where $\hat{H}$ estimates

$$E\left[\frac{(R_i^{(1)} - \mu_1)\frac{\partial \pi(X_i)}{\partial \gamma}}{\pi(X_i)} + \frac{(R_i^{(0)} - \mu_0)\frac{\partial \pi(X_i)}{\partial \gamma}}{\{1-\pi(X_i)\}}\right]$$

and $\hat{E}(S_\gamma S_\gamma^T)$ estimates $E(S_\gamma S_\gamma^T)$. Formulas for $\hat{E}(S_\gamma S_\gamma^T)$ and $\hat{H}$ are given at the end of Section 4.

By employing standard martingale computations (Fleming and Harrington, 1991), the variance of (20) will equal

$$E \int_0^L \frac{A_i}{K_1^2(u)} \times \left[\frac{(R_i^{(1)} - \mu_1)}{\pi(X_i)} - \frac{E\left[(R^{(1)} - \mu_1) I(T^{(1)} \geq u)\right]}{E\left[AI(T^{(1)} \geq u)\right]}\right]^2 \times \lambda_1^c(u) Y_i(u) du$$

$$= \int_0^L \frac{\lambda_1^c(u)}{K_1(u)} \times E\left[\left[\frac{(R_i^{(1)} - \mu_1)}{\pi(X_i)} - \frac{E\left[(R^{(1)} - \mu_1) I(T^{(1)} \geq u)\right]}{E\left[AI(T^{(1)} \geq u)\right]}\right]^2 \times A_i I(T_i^{(1)} \geq u)\right] du.$$

Consistent estimators for the variances of (20) and (21) are given by

$$\sum_{i=1}^n \frac{A_i(1-\Delta_i)\left\{\hat{G}_1(U_i)\right\}}{\left\{\sum_{l=1}^n A_l Y_l(U_i)\right\} \hat{K}_1(U_i)}$$

and

$$\sum_{i=0}^n \frac{(1-A_i)(1-\Delta_i)\left\{\hat{G}_0(U_i)\right\}}{\left\{\sum_{l=1}^n (1-A_l) Y_l(U_i)\right\} \hat{K}_0(U_i)},$$

where $\hat{G}_1(U_i)$ and $\hat{G}_0(U_i)$ are defined at the end of Section 4.