Published in final edited form as:

Biometrics. 2012 September ; 68(3): 707–716. doi:10.1111/j.1541-0420.2012.01754.x.

Estimating Propensity Scores and Causal Survival Functions Using Prevalent Survival Data

Yu-Jen Cheng^{1,*} and Mei-Cheng Wang^{2,**}

¹Institute of Statistics, National Tsing Hua University, Hsin-Chu 300, Taiwan

²Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street E3614, Baltimore, Maryland 21205, U.S.A.

Summary

This article develops semiparametric approaches for estimation of propensity scores and causal survival functions from prevalent survival data. The analytical problem arises when the prevalent sampling is adopted for collecting failure times and, as a result, the covariates are incompletely observed due to their association with failure time. The proposed procedure for estimating propensity scores shares interesting features similar to the likelihood formulation in case-control study, but in our case it requires additional consideration in the intercept term. The result shows that the corrected propensity scores in logistic regression setting can be obtained through standard estimation procedure with specific adjustments on the intercept term. For causal estimation, two different types of missing sources are encountered in our model: one can be explained by potential outcome framework; the other is caused by the prevalent sampling scheme. Statistical analysis without adjusting bias from both sources of missingness will lead to biased results in causal inference. The proposed methods were partly motivated by and applied to the Surveillance, Epidemiology, and End Results (SEER)-Medicare linked data for women diagnosed with breast cancer.

Keywords

Case-control study; Prevalent sampling; Propensity scores

1. Introduction

Suppose a disease is nonrecurrent and a failure event is the absorbing state of the disease; for example, the disease is a certain cancer or genetic disease, and the failure event is death. When studying the natural history of the disease, an incident cohort is defined as a group of subjects whose initial events are randomly sampled from a predetermined calendar time interval. The subjects are followed for detecting the occurrence of the failure event until loss to follow-up or end-of-study. When failure times are long, the incident cohort design is inefficient for natural history studies because it usually takes a long follow-up time to observe enough failure events. In contrast, a prevalent sampling design which draws subjects from a diseased population is more focused and practical in real studies. Prevalent sampling is frequently used in biological and epidemiological studies, cancer survey, and

many fields. The prevalent sample is formed by subjects whose initial event had occurred but have not experienced the failure event at the time of recruitment. When prevalent sampling is employed, it is well known that the observed failure times tend to be longer than those arising from the population survival distribution (Brookmeyer and Gail, 1987; Wang, 1991).

In a prevalent cohort study, it is typical that covariates information are collected along with failure times from each patient. In conventional regression analysis, the covariates are usually ancillary statistics if their distribution does not involve parameters in the regression model. Moreover, joint or conditional distribution among covariates can be nonparametrically estimated without specifying any particular form for outcome and covariates. However, these can not be done if patients are selected on the basis of the outcome value (Liang and Qin, 2000). Such a selection scheme includes prevalent sampling since patients tend to be observed with longer failure times. Under these circumstances, Wang (1996) proposed a weighting risk set approach to estimate the effects of covariates on outcome in Cox Model without censoring cases and extended in the presence of censoring by Qin and Shen (2010). Recently, using prevalent data to estimate joint covariates distribution has been addressed by Bergeron, Asgharian, and Wolfson (2008). In this article, the probability of treatment status given on age, gender, and other clinic characteristics is the object of interest, namely as propensity score in causal inference. The proposed estimating procedure is analogous to a case-control study with inference on a prospective likelihood but the selection probability is allowed to vary with covariates and identified from data. More precisely, the result shows that, if the propensity scores can be identified through a logistic regression model, the corrected propensity scores can then be obtained through standard estimation procedure with specific adjustments on the intercept term (Prentice and Pyke, 1979; Qin, 1998).

In observational studies the propensity score, defined as the conditional probability of being treated given the covariates, is commonly used to adjust for confounding. The propensity scores are usually unknown and need to be estimated from a parametric model. The estimated propensity scores can be used to balance covariates either by stratification or matching (Rubin, 1997; Dehejia and Wahba, 1999). Or alternatively, the inverse propensity scores can be used to adjust for confounding (Rosenbaum, 1987; Anstrom and Tsiatis, 2001; Tan, 2006). Nevertheless, these approaches are not directly applicable to prevalent survival data. When prevalent sampling is employed, the distribution of observed covariates is biased and further adjustment is needed in analytical methods.

This article has the following organization. Section 2 introduces notations and model framework. In Section 3, a new approach is proposed for estimating the propensity scores under stationary and nonstationary truncation assumptions. The proposed estimating procedure is partially similar to the likelihood structure for case-control data when making connections to inference for prospective study (Prentice and Pyke, 1979). In Section 4, based on corrected propensity scores, we propose a stratification procedure and a nonparametric smoothing approach to estimate causal survival functions. In Section 5, a simulation study is presented to examine the appropriateness and efficiencies of the proposed methods. In Section 6, we apply our methods to SEER-Medicare data for women diagnosed with breast cancer. Discussion is presented in Section 7.

2. Model and assumptions

For each subject, let T^* , W^* , A^* , and $X^* = (X_1^*, \dots, X_q^*)^T$ be the failure time, truncation time, treatment status, and a $q \times 1$ vector of covariates, respectively. The variables with superscript asterisk represent the population variables; variables without asterisk denote

variables subject to truncation, that is, $\text{pr}(T, W, A, X) = \text{pr}(T^*, W^*, A^*, X^*/T^* = W^*)$. For subjects in the disease prevalent population, define C as the residual censoring time, or equivalently, time from recruitment to censoring. Let $Y = \min(T, W + C)$ and $\delta = I(T < W + C)$. The typical observations from an observational study with prevalent sampling include n independent random vectors $\{Y_i, \delta_i, W_i, A_i, X_i; i = 1, \dots, n\}$. In this article, we develop methods for estimation of the propensity scores using prevalent data.

2.1 Assumptions and Likelihood

Prevalent survival data are commonly encountered in biomedical studies. For example, based on information from SEER-Medicare data, T^* is the time from diagnosis with breast cancer to death for a patient without ($A^* = 0$) or with ($A^* = 1$) radiation therapy and W^* is the time from diagnosing with breast cancer to recruitment. For each subject, Y is obtained only when $T^* = W^*$. Throughout this article, the following assumption is assumed:

- (A1) *Independent censoring and truncation assumption:* (i) T^* is conditionally independent of W^* given (A^*, X^*) , i.e., $\text{pr}(T^*, W^*/A^*, X^*) = \text{pr}(T^*/A^*, X^*)\text{pr}(W^*/A^*, X^*)$. (ii) Conditioning on (A^*, X^*) and $T^* = W^*$, the residual censoring time C (from recruitment to censoring) is independent of the residual failure time $T - W$ (from recruitment to failure event) for subjects in the disease prevalent population.

Note that Assumption (A1) indicates independent censoring assumption only on the disease prevalent population.

For simplicity of notation, let $G_{W^*, A^*, X^*}(w, a, x) = \text{pr}(W^* = w, A^* = a, X^* = x)$, $G_{W^*/A^*, X^*}(w/a, x) = \text{pr}(W^* = w/a, A^* = a, X^* = x)$, $G_{A^*, X^*}(a, x) \equiv G_{W^*, A^*, X^*}(\infty, a, x) = \text{pr}(W^* = \infty, A^* = a, X^* = x)$, and $G_{X^*}(x) \equiv \text{pr}(X^* = x, a = 0, 1)$. To avoid identifiability problems, a technical assumption is assumed: the minimum of support of truncation time must be less than minimum of support of failure time, and the maximum of support of truncation time must be less than the maximum of support of failure time. When assumptions (A1) holds, the full likelihood, L_F for observed data $\{y_i, \delta_i, w_i, a_i, x_i; i = 1, \dots, n\}$ can be further decomposed into two parts: the marginal likelihood of observed $((w_1, a_1, x_1), \dots, (w_n, a_n, x_n))$,

$$L_M = \prod_{i=1}^n \frac{S(w_i|a_i, x_i)dG_{W^*, A^*, X^*}(w_i, a_i, x_i)}{\int S(w_i|a_i, x_i)dG_{W^*, A^*, X^*}(w_i, a_i, x_i)}, \quad (1)$$

and the conditional likelihood of $([y_1, \delta_1], \dots, [y_n, \delta_n])$ given $([w_1, a_1, x_1], \dots, [w_n, a_n, x_n])$,

$$L_C \propto \prod_{i=1}^n \left(\lambda_{T^*}[y_i|a_i, x_i]^{\delta_i} \frac{S[y_i|a_i, x_i]}{S[w_i|a_i, x_i]} I[y_i \geq w_i] \right), \quad (2)$$

where $\lambda_{T^*}(u/A^*, X^*)$ and $S(u/A^*, X^*)$ are the hazard and survival function of T^* at u conditional on treatment assignment A^* and covariates X^* , respectively. Based on this decomposition, it explicitly shows that $S(w/a, x)$ plays the role as a selection-biased function when data are observed subject to prevalent sampling. It will be seen later that this selection-biased function can be estimated from the conditional likelihood (2).

2.2 The Outcome Model

Consider the proportional hazard model (Cox, 1972) that the hazard function of the failure time T^* , given covariates A^* and X^* , is

$$\lambda_{T^*}(u|A^*=a, X^*=x)=\lambda_0(u)\exp (a\gamma+x^T\beta). \quad (3)$$

where $\lambda_0(u)$ is an unspecified function, γ is the log hazard ratio for the effect of radiation therapy on survival time, and $\beta = (\beta_1, \dots, \beta_q)$ is a $q \times 1$ vector of regression parameters.

By substituting the hazard model (3) into the full likelihood $L_F L_C$ can be further expressed as the product of a partial likelihood, L_P and residual likelihood L_R , where

$$L_P = \prod_{(i)} \frac{\exp (a_{(i)}\gamma + x_{(i)}^T\beta)}{\sum_{j \in R(y_{(i)})} \exp (a_j\gamma + x_j^T\beta)}$$

and $y_{(1)}, \dots, y_{(k)}$ are distinct ordered uncensored failure times, and $R(y_{(i)}) = \{j: w_j \leq y_{(i)}\}$, $j=1, \dots, n\}$. The estimators $\hat{\gamma}$ and $\hat{\beta}$ of γ and β can be obtained by maximizing the partial likelihood function L_P . The Breslow estimator (Breslow, 1972) for the cumulative hazard function $\Lambda_0(u) = \int_0^u \lambda_0(t)dt$, is

$$\hat{\Lambda}_0(u) = \sum_{y_{(i)} \leq u} \frac{d_{(i)}}{\sum_{j \in R(y_{(i)})} \exp (a_j\hat{\gamma} + x_j^T\hat{\beta})},$$

where $d_{(i)} = \sum_{j=1}^n I(y_j = y_{(i)}, \delta_j = 1)$. Thus, the function $S(u/a, x)$ can be estimated through the Cox model with the plug-in estimators, that is, $\hat{S}(u/a, x) = \exp\{-\hat{\Lambda}_0(u) \exp(a\hat{\gamma} + x^T\hat{\beta})\}$.

Further, the marginal likelihood (1) can be treated as a multinomial distribution with selection-biased probabilities. By plugging the consistent estimator of selection-biased function, $\hat{S}(w/a, x)$, the distribution of (W^*, A^*, X^*) can be estimated by the inverse-weighting estimator

$$\hat{G}_{W^*, A^*, X^*}(w, a, x) = \left(\sum_{i=1}^n \frac{1}{\hat{S}(w_i|a_i, x_i)} \right)^{-1} \times \sum_{i=1}^n \frac{I(w_i \leq w, a_i = a, x_i \leq x)}{\hat{S}(w_i|a_i, x_i)}.$$

Therefore, the marginal distribution of X^* are estimated as

$$\hat{G}_{X^*}(x) = \left(\sum_{i=1}^n \frac{1}{\hat{S}(w_i|a_i, x_i)} \right)^{-1} \sum_{i=1}^n \frac{I(x_i \leq x)}{\hat{S}(w_i|a_i, x_i)}.$$

Note that this result shows that estimating $G_{X^*}(x)$ by the empirical distribution from $\{x_i\}$ is inappropriate since the distribution of observed covariates is also biased when prevalent sampling scheme is adopted to collect failure times.

3. A Semiparametric Approach

The propensity score is defined as the probability of being treated conditional on individual's covariate values: $\alpha(x) = \text{pr}(A^* = 1|X^* = x)$.

It is indicated in Section 2.2 that the covariates are observed subject to sampling bias when prevalent sampling scheme is applied to collect failure time data. Thus, it is inappropriate to use the conventional approach to estimate $\epsilon(x)$. In this section we present a new approach to correct the bias.

3.1 Estimation of Propensity Score

The propensity score is assumed to follow a logit model:

$$\text{pr}(A^*=1|X^*=x) = \frac{\exp(\theta_0 + x^T \theta)}{1 + \exp(\theta_0 + x^T \theta)}, \quad (4)$$

where θ_0 is the intercept and $\theta = (\theta_1, \dots, \theta_q)$ is a $q \times 1$ vector of parameters. Regardless that covariates are more likely to be observed with longer failure times or shorter truncation times as shown in Section 2.2, the explicit sampling bias from observed covariates is still unclear and needs to be explored. Specifically, the conditional distribution of observed treatment assignment A given other observed covariates X should be identified. To explicitly identify this dependency, we explore the relationship between observed (A, X) and (A^*, X^*) by integrating out the joint distribution of (W, A, X) over W :

$$\text{pr}(A=a, X=x) \propto \left(\int S(w|a, x) dG_{W^*|A^*, X^*}(w|a, x) \right) \times \text{pr}(A^*=a, X^*=x).$$

The integral $\int S(w|a, x) dG_{W^*|A^*, X^*}(w|a, x)$ can be rewritten as $\text{pr}(T^* \leq W^*/a, x)$, which is the proportion of untruncated data given A^* and X^* . From the expression above, it explicitly shows that the joint distribution of observed (A, X) is related to the joint distribution of (A^*, X^*) through a selection bias function $\text{pr}(T^* \leq W^*/a, x)$. Furthermore, by applying Bayes' rule, the conditional distribution of A given X can be shown as

$$\text{pr}(A=1|X=x) = \frac{\text{pr}(A=1, X=x)}{\text{pr}(A=0, X=x) + \text{pr}(A=1, X=x)} = \frac{\exp(\log[\alpha(x; \Lambda_0, \gamma, \beta)] + \theta_0 + x^T \theta)}{1 + \exp(\log[\alpha(x; \Lambda_0, \gamma, \beta)] + \theta_0 + x^T \theta)}, \quad (5)$$

where $\alpha(x; \Lambda_0, \gamma, \beta) = \frac{\text{pr}(T^* \geq W^*|A^*=1, X^*=x)}{\text{pr}(T^* \geq W^*|A^*=0, X^*=x)}$ is a ratio of untruncated probability conditional on treatment assignment and other covariates. In nonparametric settings, without conditioning on treatment assignment and covariates, the untruncated probability was previously studied by several authors in absence of censoring (Woodroffe, 1985; Chao, 1987; He and Yang, 1998), and other authors extended the results to allow for censoring (Wang, 1991; Shen, 2005). In this article, the estimation of untruncated probability conditional on treatment assignment and covariates is of interest. By incorporating the information of the outcome model and further assuming that truncation times only depends on treatment assignment, the untruncated probability, $\text{pr}(T^* \leq W^*/a, x)$, is estimated by

$$\sum_{i=1}^n \widehat{S}(w_i|a, x) d\widehat{G}_{W^*|A^*, X^*}(w_i|a, x), \text{ where}$$

$$\widehat{G}_{W^*|A^*, X^*}(w|a, x) = \widehat{G}_{W^*|A^*}(w|a) = \left(\sum_{i=1}^n \frac{I(a_i=a)}{\widehat{S}(w_i|a_i, x_i)} \right)^{-1} \sum_{i=1}^n \frac{I(a_i=a, w_i \leq w)}{\widehat{S}(w_i|a_i, x_i)}, \text{ an analogue to}$$

nonparametric estimate of $G_{W^*}(w)$ (Wang, 1991). Therefore, $\log\{\alpha(x; \Lambda_0, \gamma, \beta)\}$ can be estimated as

$$\log(\hat{\alpha}(x; \hat{\Lambda}_0, \hat{\gamma}, \hat{\beta})) = \log \frac{\left(\sum_{i=1}^n \frac{I[a_i=0]}{\hat{S}[w_i|a_i, x_i]} \right) \left(\sum_{i=1}^n \frac{\hat{S}[w_i|A^*=1, X^*=x] I[a_i=1]}{\hat{S}[w_i|a_i, x_i]} \right)}{\left(\sum_{i=1}^n \frac{I[a_i=1]}{\hat{S}[w_i|a_i, x_i]} \right) \left(\sum_{i=1}^n \frac{\hat{S}[w_i|A^*=0, X^*=x] I[a_i=0]}{\hat{S}[w_i|a_i, x_i]} \right)}.$$

So, by substituting $\log\{\hat{\alpha}(x; \hat{\Lambda}_0, \hat{\gamma}, \hat{\beta})\}$ into equation (5), the parameters θ_0 and θ can be estimated by the original logistic regression model but adjusted for the “offset” term, $\log\{\hat{\alpha}(x; \hat{\Lambda}_0, \hat{\gamma}, \hat{\beta})\}$. As the result, for each subject, the propensity score can be estimated as

$$\hat{e}(x) = \frac{\exp(\hat{\theta}_0 + x^T \hat{\theta})}{1 + \exp(\hat{\theta}_0 + x^T \hat{\theta})}.$$

3.2 Link to Case-Control Study

As an interesting finding, the probability of A given X again possesses a logistic form and the odds ratio stays as the multiplicative parameter in the logistic transformation. For easy distinction, in the rest of our article, we refer model (5) as a prevalent model and model (4) as an incident model. Models (4) and (5) are almost equivalent except for the “offset” term, $\log\{\alpha(x; \Lambda_0, \gamma, \beta)\}$. Using this specific characteristic, we can then estimate θ_0 and θ from the prevalent data using standard software for logistic regression by adding the “offset” term, $\log\{\alpha(x; \Lambda_0, \gamma, \beta)\}$.

Note that a similar phenomenon was found in case-control studies and has been addressed (Prentice and Pyke, 1979). In a case-control study, each case ($A = 1$) or control ($A = 0$) is sampled from a case or control population with a selection probability only depending on $A^* = 1$ or 0, say q_1 and q_0 . Then, if (A, X) is sampled from a case-control study, the relationship between (A, X) and (A^*, X^*) is

$$\text{pr}(A=1|X=x) = \frac{\exp(\log[q_1/q_0] + \theta_0 + x^T \theta)}{1 + \exp(\log[q_1/q_0] + \theta_0 + x^T \theta)}. \quad (6)$$

In contrast to the expression (6), $\text{pr}(T^* \geq W^*|A^*=1, X^*=x)$ plays a role as a selection probability for each case and control. However, in such a connection, unlike the case-control study, the selection probability in our case is allowed to vary with covariates and can be identified from data.

Qin (1998) considered a case-control study with biased sampling setting which can be also linked to our prevalent sampling model. Using Bayes' rule again, the conditional distribution of X given A can be shown as

$$\text{pr}(X=x|A=a) = \left(\frac{\text{pr}[T^* \geq W^*|A^*=a, X^*=x]}{\text{pr}[T^* \geq W^*|A^*=a]} \right) \times \left(\frac{\text{pr}[A^*=a|X^*=x]}{\text{pr}[A^*=a]} \right) \text{pr}(X^*=x)$$

and the ratio between case and control is

$$\frac{\text{pr}(X=x|A=1)}{\text{pr}(X=x|A=0)} = \left(\frac{\text{pr}[T^* \geq W^*|A^*=1, X^*=x] \text{pr}[T^* \geq W^*|A^*=0]}{\text{pr}[T^* \geq W^*|A^*=1] \text{pr}[T^* \geq W^*|A^*=0, X^*=x]} \right) \times \left(\frac{\text{pr}[A^*=0]}{\text{pr}[A^*=1]} \right) \times \exp(\theta_0 + x^T \theta).$$

Thus, if the selection probability $\text{pr}(T^* = W^*/A^* = 1, X^* = x)$ only depends on A^* , the ratio becomes

$$\frac{\text{pr}(X=x|A=1)}{\text{pr}(X=x|A=0)} = \left(\frac{\text{pr}[A^*=0]}{\text{pr}[A^*=1]} \right) \exp(\theta_0 + x^T \theta),$$

which is the same formula used by Qin (1998).

If we restrict our attention on a stationary truncation assumption and only allow for prevalent cases, then equation (5) is reduced as the one in Begg and Gray (1987), i.e.,

$$\text{pr}(T^* \geq W^* | A^*=1, X^*=x) \propto \mu_{1,x}, \text{pr}(T^* \geq W^* | A^*=0, X^*=x) = q_0, \text{logit pr}(A=1|X=x) \propto \theta_0 + x^T \theta + \log \mu_{1,x},$$

where $\mu_{a,x} = \int S(w/a, x) dw$, $a = 0$ or 1 . More details for the formula under a stationary truncation assumption will be discussed in the next subsection.

3.3 Length-Bias Sampling

In this subsection, we consider the special case when the incidence of breast cancer occurs with a constant rate over time. This phenomenon has been discussed by several authors (Vardi, 1982; Wang, 1991; Asgharian, M'Lan, and Wolfson, 2002; Luo and Tsai, 2009; Shen, Ning, and Qin, 2009) and is referred as a stationary truncation assumption. The observed y_i in such a condition is well known to be length-biased. Under stationary truncation assumption, one can further simplify the ratio of untruncated probability, $\alpha(x; \Lambda_0, \gamma, \beta)$ as

$\frac{\mu_{1,x}}{\mu_{0,x}}$. Hence, $\alpha(x; \Lambda_0, \gamma, \beta)$ can be estimated as

$$\widehat{\alpha}^{LB}(x; \widehat{\Lambda}_0, \widehat{\gamma}, \widehat{\beta}) = \frac{\sum_{j=1}^n \exp(-\widehat{\Lambda}_0[w_j] \exp[\widehat{\gamma} + x^T \widehat{\beta}])}{\sum_{j=1}^n \exp(-\widehat{\Lambda}_0[w_j] \exp[x^T \widehat{\beta}])}.$$

Similarly, as done in Section 3.1, one can obtain a more efficient estimator of propensity

$$\widehat{e}^{LB}(x) = \frac{\exp(\widehat{\theta}_0^{LB} + x^T \widehat{\theta}^{LB})}{1 + \exp(\widehat{\theta}_0^{LB} + x^T \widehat{\theta}^{LB})}.$$

4. Causal Estimation using Prevalent Data

In observational studies, investigators frequently have limited or no control over the treatment assignment. Therefore, one can get a biased result in estimating the causal treatment effect due to distributional differences on observed covariates between treated and nontreated groups. In these cases, a propensity score approach can be used to reduce bias (Rosenbaum and Rubin, 1983).

Let T_0^* or T_1^* be the potential outcomes that is observed if the subject received treatment “0” or “1”. Define T^* as either T_0^* or T_1^* , depending on whichever is observed. Using definitions introduced in Section 2, the observed Y is obtained from two step procedures:

$T^* = (1 - A^*)T_0^* + A^*T_1^*$ and $Y = \min(T, W + C)$. Thus, the observed Y contains two

nonignorable sources of missingness: (1) The two components (T_0^*, T_1^*) , are not simultaneously observable since one necessarily excludes the other, so the concept of potential quantities necessarily imply missing information and (2) Y is obtained only when $T^* = W^*$. To address the complexity of data from these two sources of missingness, we consider the following assumptions:

- (A2) *Stable Unit Treatment Value Assumption (SUTVA)*: Response of the i th subject should not be affected by responses of other individuals (Noninterference). Treatment A^* could be assigned by a number of ways, but all those ways should lead to the same outcome (Consistency).
- (A3) *Strongly ignorable treatment assignment*: the treatment assignment A^* is independent of potential outcomes, (T_0^*, T_1^*) , given covariates X^* , i.e., $\text{pr}(A^*|X^*, T_0^*, T_1^*) = \text{pr}(A^*|X^*)$ and $0 < \text{pr}(A^* = a|X^*) < 1$.

The objective here is to estimate the causal survival functions based on prevalent sample, where the causal survival functions, $S_0(t)$ and $S_1(t)$, are defined as

$S_0(t) = \text{pr}(T_0^* > t)$ and $S_1(t) = \text{pr}(T_1^* > t)$, respectively. With given propensity scores, Rosenbaum and Rubin (1983) showed that A^* is independent of (T_0^*, T_1^*) if assumption (A3) holds.

Applying their inferential result to our case, the causal survival function

$S_a(u) = E_{\epsilon(X^*)} \{ \text{pr}(T_a^* \geq u | \epsilon(X^*)) \}$ is equivalent to $E_{\epsilon(X^*)} \{ \text{pr}(T^* \geq u | A^* = a, \epsilon(X^*)) \}$. Therefore, one can use stratification or univariate nonparametric regression methods to estimate $E_{\epsilon(X^*)} \{ \text{pr}(T^* \geq u | A^* = a, \epsilon(X^*)) \}$ by substituting $\epsilon(x)$ with $\hat{\epsilon}(x)$ (Rosenbaum and Rubin, 1983; Heckman, Ichimura, and Todd, 1997). These estimation steps can be considered as model-based standardization procedures for estimation of the population-average quantities (Lane and Nelder, 1982; Greenland, 2004). In our case, however, such standardization procedures alone are insufficient for causal estimation and some further modifications are needed to correct bias from the prevalent survival data. More detailed discussions are provided in the next two subsections.

4.1 Stratification

With stratification, all subjects are separated into M strata on the basis of each subject's propensity score. After checking bias-corrected covariates within each stratum are balanced between treatment and nontreatment groups, $S_a(u)$ can be approximated as

$$S_a(u) \approx \sum_{m=1}^M S_{a,m}(u) \text{pr}(e_m),$$

where $S_{a,m}(u)$ is the survival function for group $A = a$ in stratum m and e_m is a set of subjects included in stratum m . More precisely, one can implement this by dividing the interval into M strata with size equal to $1/M$, so that

$$e_m = \left(x_i : Q^{-1} \left[\frac{m-1}{M} \right] \leq \widehat{e}[x_i] < Q^{-1} \left[\frac{m}{M} \right], i=1, \dots, n, Q \text{ is the empirical distribution of } \widehat{e}[x] \right)$$

for $m = 1, \dots, M$.

Given these strata, one can estimate the survival function for group $A = a$ in stratum m as if random assignment held:

$$\widehat{S}_{a,m}(u) = \prod_{y_i \leq u, \delta_i = 1, a_i = a, x_i \in e_m} (1 - \widehat{\lambda}_{im}),$$

where $\widehat{\lambda}_{im} = \frac{\sum_{j=1}^n I(y_j = y_{(i)}, \delta_j = 1, a_j = a) I(x_j \in e_m)}{\sum_{j=1}^n I(w_j \leq y_{(i)} \leq y_j, a_j = a) I(x_j \in e_m)}$. Then we can estimate $S_a(u)$ with $\widehat{S}_{a,m}(u)$ substituted by $\widehat{S}_{a,m}(u)$ in calculation of $\text{pr}(e_m)$. However, it is inappropriate to use $1/M$ for estimating $\text{pr}(e_m)$ since subjects are observed subject to prevalent sampling bias. Because propensity score is a function of covariates, we can use the result in Section 2.2 and take the sum of the weights of subjects in each stratum, that is,

$$\widehat{\text{pr}}(e_m) = \sum_{x_j \in e_m} d\widehat{G}_{x^*}(x_j).$$

Note that it would be necessary to give each subject an appropriate weight when examining the balance of covariates within each stratum. A proper reweighting procedure is to use the inverse-probability-weighted estimate of the distribution of covariates.

4.2 Nonparametric Regression

In the stratification approach, the estimators of $S_a(u)$ can be affected by the number of strata and lead to a biased result. To avoid the bias due to subjective selection of M , we can take an alternative approach by treating the estimated propensity scores as a covariate which condenses high dimensional information and uses the reduced data set $\{y_i, \delta_i, w_i, a_i, \widehat{e}(x_i); i = 1, \dots, n\}$ to draw inference on $S_a(u)$, $a=0, 1$.

Without making any further assumption, we consider a nonparametric approach to estimate the cumulative hazard function of T_a^* given $\mathcal{A}(x)$, $H_a\{u/\mathcal{A}(x)\}$, by a kernel smoothing method (Perez and Manteiga, 1999):

$$\widehat{H}_a\{u/\mathcal{A}(x)=z\} = \frac{\sum_{j=1}^n I(a_j=a, y_j=y_{(i)}) \delta_j K\left(\frac{\widehat{e}(x_j) - z}{b_n}\right)}{\sum_{j=1}^n I(a_j=a, y_j \geq y_{(i)} \geq w_j) K\left(\frac{\widehat{e}(x_j) - z}{b_n}\right)},$$

where $K(\cdot)$ is a kernel function, b_n is a bandwidth with choice. The estimator for the conditional causal survival function of T_a^* given $\mathcal{A}(x) = z$ is then $\widehat{S}_a(u/z) = \prod_{t \leq u} [1 - d\widehat{H}_a\{t/\mathcal{A}(x) = z\}]$. The causal survival functions can therefore be estimated as a weighted average of $\widehat{S}_a(u/\widehat{e}(x_l))$ over the distribution of $\widehat{e}(x_l)$, $l = 1, \dots, n$, that is,

$$\widehat{S}_a^k(u) = \sum_{l=1}^n \left[\prod_{y_{(i)} \leq u} \left[1 - \frac{\sum_{j=1}^n I\{a_j=a, y_j=y_{(i)}\} \delta_j K\left\{\frac{\widehat{e}(x_j) - \widehat{e}(x_l)}{b_n}\right\}}{\sum_{j=1}^n I\{a_j=a, y_j \geq y_{(i)} \geq w_j\} K\left\{\frac{\widehat{e}(x_j) - \widehat{e}(x_l)}{b_n}\right\}} \right] \right] \times \widehat{\text{pr}}\{\widehat{e}(x)\}.$$

To accomplish estimation of the causal survival functions, one can estimate $\text{pr}(\mathcal{A}(x) = z)$ by $x: \hat{e}(x) = z \frac{d\hat{G}_{X^*}(x)}{dx}$. Note that:

1. In the form of our estimators the outer summation is with respect to the marginal distribution of $\mathcal{A}(X^*)$, not the conditional distribution of $\mathcal{A}(X^*)$ given $A^* = 1$ or 0 .
2. $G_{X^*}(x)$ can be thought of as a probability weighting function performed for both treated and untreated groups.

5. Simulation Study

In this section, we present four simulation studies to compare the performance of our proposed estimators in Sections 3.1 and 4. For each subject i , we generated X_{1i}^* from a Bernoulli distribution with a parameter $p = 0.6$ and X_{2i}^* from a continuous uniform(0, 1) distribution, where X_{1i}^* and X_{2i}^* are independent of each other. After generating X_{1i}^* and X_{2i}^* , A_i^* can be generated from a Bernoulli distribution with probability $\frac{\exp(\theta_0 + X_{1i}^*\theta_1 + X_{2i}^*\theta_2)}{1 + \exp(\theta_0 + X_{1i}^*\theta_1 + X_{2i}^*\theta_2)}$ and let $(\theta_0, \theta_1, \theta_2) = (1, 1, 1)$ and $(1, -1, -1)$ correspond to different treatment assignment mechanisms. The observed data vectors were generated according to the following algorithm.

- Use a constant baseline hazard function $\lambda_0(t) = \xi$, where $\xi = 0.05$.
- Conditional on X_{1i}^* and X_{2i}^* , generate potential survival times, T_{0i}^* and T_{1i}^* , with a exponential distribution with failure rate $\xi \exp(X_{1i}^*\beta_1 + X_{2i}^*\beta_2)$ and $\xi \exp(\gamma + X_{1i}^*\beta_1 + X_{2i}^*\beta_2)$, respectively, where $\gamma = 1$ and $(\beta_1, \beta_2) = (1, 1)$.
- The survival time is obtained by $T_i^* = A_i^*T_{1i}^* + (1 - A_i^*)T_{0i}^*$.
- Generate truncation time, W_i^* from an uniform distribution on the interval $(0, B)$, where $B = 3$ and 5 are corresponding to different degrees of truncation.
- The residual censoring time, C_i , was generated from an Uniform(0,100) variable.
- The data are only observed when $T_i^* \geq W_i^*$. The observed data vector for subject i is $(Y_i, \delta_i, W_i, A_i, X_i)_{i=1}^n$, where $Y_i = \min(T_i^*, W_i^* + C_i)$, $\delta_i = I(T_i^* < W_i^* + C_i)$.

Four simulation scenarios, with $(p, B, \theta_1, \theta_2) = (0.6, 5, 1, 1), (0.6, 3, 1, 1), (0.6, 5, -1, -1)$, and $(0.6, 3, -1, -1)$, are considered in this study, corresponding to truncation and censoring rate (55%, 4%), (41%, 4%), (44%, 6%), and (31%, 6%), respectively. Using this algorithm, we generated $M = 500$ data sets with 1,000 subjects in each scenario, after truncation, the average sample sizes are around 450, 590, 560, and 690, respectively.

Figure 1 provides a comparison between Na and SBC methods for estimating the parameters in the propensity score model, where Na denotes the naïve approach without accounting the truncation effect and SBC denotes the approaches as described in Section 3.1. It is clear to see that Na estimator is underestimated for parameters θ_0, θ_1 , and θ_2 through all scenarios. Moreover, the magnitude of bias also depends on the degree of truncation, for example, this situation becomes more severe in scenario 1 than in scenario 2, and more severe in scenario 3 than in scenario 4. On the other hand, the proposed SBC estimators consistently perform well in all simulation scenarios. We also considered the length-bias estimators as described in Section 3.3. The performance of the length-bias estimators appears similar to SBC methods but shows slight improvement in efficiency.

Figure 2 presents boxplots of estimated 5-year survival rates for both treated and control groups. For convenience, we use S and K to denote the proposed stratification and kernel smoothing approaches as described in Sections 4.1 and 4.2, respectively. For stratification analysis, patients were divided into five strata of approximately equal size on the basis of propensity scores, in contrast with the kernel smoothing analysis with bandwidth $b_n = 0.00005$. These approaches are also compared to the stratification approach (denotes by NS) (Rubin, 1997) and the kernel smoothing method (denotes by NK) (Heckman et al., 1997) without accounting for truncation effect. It can be found that our estimators consistently perform well in all simulation scenarios. In contrast, both NS and NK approaches tend to overestimate the survival rates due to ignorance of truncation effect.

6. Data Application

The proposed methods are applied to Surveillance, Epidemiology, and End Results (SEER)-Medicare data (Warren et al., 2002), the linkage of cancer registry data from the National Cancer Institute (NCI) and Medicare claims from the Centers for Medicare and Medicaid Services (CMS). The SEER data consist of information from cancer patients diagnosed between 1973 and 2007. The medicare data contain treatment and medical cost information between 1986 and 2007. The survival time of interest is the time from diagnosis to death. The linkage of the two data sets creates an incident cohort as well as a prevalent cohort of underlying population. Combining the two cohorts, survival data from the combined cohort can be thought of as left truncated since Medicare claims are only available for patients still surviving in 1986. The scientific question of interest in our analysis is the necessity of additional breast radiation therapy for older women with breast-conserving surgery (lumpectomy, with and without dissection of axillary lymph nodes) in the early stage of breast cancer (Buchholz, 2009).

To address the question, we used the SEER-Medicare database to identify women aged over 65 years with Stage 0, 1, and 2. The stage grouping is defined by the American Joint Committee (AJC) (Kinne, 1991). To better examine the truncation effect, we only consider patients diagnosed with breast cancer before 1986. In this prevalent data set, two thousand women diagnosed with breast cancer were included and a total of 969 patients were treated with radiation therapy. Initially, we started the analysis by focusing on those confounding factors considered by Rubin (1997), which include age, race (2 levels), marital status (2 levels), stage (3 levels), laterality (2 levels) and geographical registry (9 levels).

The proposed estimates of the difference (treatment-control) in survival at 5 years are -0.0173 (0.032) and -0.0168 (0.034) along with 95% bootstrap confidence interval $(-0.078, 0.045)$ and $(-0.084, 0.051)$ for stratification and kernel smoothing, respectively, where the values in the bracket are bootstrap estimates of standard error. In Figure 3, it gives a plot of the distribution of 5-year survival rates of the proposed estimators from the bootstrap sampling (stratification: S and kernel smoothing: K) and their corresponding naïve estimators, NS and NK, respectively for treatment (with radiation therapy) and control (without radiation therapy) groups. For stratification analysis, in this data application, patients were divided into five strata of approximately equal size on the basis of the individual propensity score, where a bandwidth $b_n = 0.00005$ is used in the kernel smoothing analysis. It can be found that both naïve estimators for 5-year survival rate tend to be higher than our estimator in both cases, with and without radiation therapy. Furthermore, the proposed stratification estimator performs similarly to the proposed kernel smoothing estimator. More analyzes of the data application, like the number of subjects in each stratum, estimated survival curves, and the information about truncation times, can be found on the Web Appendix as supplementary materials.

7. Discussion

In this article, we proposed an approach to estimate propensity scores and causal survival functions from prevalent survival data. The complexity of the research problem arises when observational bias and prevalent sampling bias are both present in a study. In our formulation of distributional bias, it is seen that the empirical distribution of observed data can be expressed as a weighted distribution of the population distribution with proper weight. Statistical issues of similar features have recently caught attentions in other areas; for example, Heckman (2008) pointed out the importance of econometric causality in exploration of the difference between sampling and population distributions. In epidemiology, this issue is related to causal effect estimation in case-control study (Joffe and Rosenbaum, 1999; Mansson et al., 2007; Rose and van der Laan, 2009), in which the research interest is causal odds ratio but the data are collected subject to biased sampling. In our case, with truncation effect on survival data, covariates can only be partially observed due to their association with failure time and therefore form an additional source of sampling bias.

In our estimation of propensity scores, the proposed procedure shares interesting features which are similar to the likelihood formulation in case-control studies, but it requires additional consideration to adjust for bias. In our logistic regression model, the bias can be adjusted by adding the “offset” term as seen in Section 3, and the “offset” term has an explicit expression as the ratio of untruncated probabilities which can be estimated from the failure time model. Moreover, the proposed estimating procedure does not necessarily rely on the Cox regression model and can be extended to other failure time models such as the accelerated failure time model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by the Taiwan National Science Council NSC99-2118- M-007-003- (to first author) and National Institute of Health grant R01AI078835 (to second author). The authors thank the associate editor and two referees for very helpful and constructive comments.

References

- Anstrom KJ, Tsiatis AA. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics*. 2001; 57:1207–1218. [PubMed: 11764262]
- Asgharian M, M’Lan CE, Wolfson DB. Length-biased sampling with right censoring: An unconditional approach. *Journal of the American Statistical Association*. 2002; 97:202–209.
- Begg C, Gray R. Methodology for case-control studies with prevalent cases. *Biometrika*. 1987; 74:191–195.
- Bergeron P, Asgharian M, Wolfson D. Covariate bias induced by length-bias sampling of failure times. *Journal of the American Statistical Association*. 2008; 103:737–742.
- Breslow NE. Contribution to the discussion of the paper by D.R. Cox. *Journal of the Royal Statistical Society Series B*. 1972; 34:216–217.
- Brookmeyer R, Gail MH. Biases in prevalent cohorts. *Biometrics*. 1987; 43:739–749. [PubMed: 3427161]
- Buchholz TA. Radiation therapy for early-stage breast cancer after breast-conserving surgery. *The New England Journal of Medicine*. 2009; 360:63–70. [PubMed: 19118305]
- Chao MT. Influence curves for randomly truncated data. *Biometrika*. 1987; 74:426–429.

- Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B*. 1972; 34:187–220.
- Dehejia RH, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*. 1999; 94:1053–1062.
- Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology*. 2004; 160:301–305. [PubMed: 15286014]
- He S, Yang GL. Estimation of the truncation probability in the random truncation model. *The Annals of Statistics*. 1998; 26:1011–1028.
- Heckman J, Ichimura H, Todd P. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*. 1997; 64:605–654.
- Heckman JJ. Econometric causality. *International Statistical Review*. 2008; 76:1–27.
- Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *American Journal of Epidemiology*. 1999; 150:327–333. [PubMed: 10453808]
- Kinne DW. Staging and follow-up of breast cancer patients. *Cancer*. 1991; 67(4):1196–1198. [PubMed: 1991279]
- Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics*. 1982; 38:613–621. [PubMed: 7171691]
- Liang KY, Qin J. Regression analysis under non-standard situations: A pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B*. 2000; 62:773–786.
- Luo XD, Tsai WY. Nonparametric estimation for right-censored length-biased data: A pseudo partial likelihood approach. *Biometrika*. 2009; 96:873–886.
- Mansson R, Joffe MM, Sun W, Heanessy S. On the estimation and use of propensity scores in case-control and casecohort studies. *American Journal of Epidemiology*. 2007; 166:332–339. [PubMed: 17504780]
- Perez CI, Manteiga WG. Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *Journal of Nonparametric Statistics*. 1999; 10:213–244.
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
- Qin J. Inferences for case-control and semiparametric twosample density ratio models. *Biometrika*. 1998; 85:619–630.
- Qin J, Shen Y. Statistical methods for analyzing rightcensored length-biased data under cox model. *Biometrics*. 2010; 66:382–392. [PubMed: 19522872]
- Rose S, van der Laan MJ. Why match? Investigating matched case-control study designs with causal effect estimation. *The International Journal of Biostatistics*. 2009; 5(1) Article 1.
- Rosenbaum P. Model-based direct adjustment. *Journal of the American Statistical Association*. 1987; 82:387–394.
- Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*. 1997; 127:757–763. [PubMed: 9382394]
- Shen PS. Estimation of the truncation probability with lefttruncated and right censored data. *Journal of Nonparametric Statistics*. 2005; 17:957–969.
- Shen Y, Ning J, Qin J. Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association*. 2009; 104:1192–1202. [PubMed: 21057599]
- Tan Z. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*. 2006; 101:1619–1637.
- Vardi Y. Nonparametric estimation in the presence of length bias. *Annals of Statistics*. 1982; 10:616–620.
- Wang MC. Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*. 1991; 86:130–143.
- Wang MC. Hazards regression analysis for length-biased data. *Biometrika*. 1996; 83:343–354.

- Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the seer-medicare data: Content, research applications, and generalizability to the united states elderly population. *Med Care*. 2002; 40:3–18.
- Woodroffe M. Estimating a distribution function with truncated data. *The Annals of Statistics*. 1985; 13:163–177.

Figure 1.

Boxplots of estimated coefficients in propensity scores model in four different simulation scenarios. Na denotes the naïve approach without accounting for the truncation effect and SBC denotes the approaches as described in Section 3.1. Dash lines denote the true values.

Figure 2.

Boxplots of estimated 5-year survival rates for both treated (with radiation therapy) and control group (without radiation therapy). NS: naïve stratification approach without accounting the truncation effect; S: proposed stratification approach as described in Section 4.1; NK: naïve kernel smoothing approach without accounting the truncation effect; K: proposed kernel smoothing approach as described in Section 4.2. Dash lines denote the true values.

Figure 3.

Boxplots of the distribution of estimated 5-year survival rates from the bootstrap sampling for both treated (with radiation therapy) and control group (without radiation therapy). NS: naïve stratification approach without accounting the truncation effect; S: proposed stratification approach; NK: naïve kernel smoothing approach without accounting the truncation effect; K: proposed kernel smoothing approach.