HHS Public Access

Author manuscript

Clin Trials. Author manuscript; available in PMC 2016 August 01.

Published in final edited form as:

Clin Trials. 2015 August ; 12(4): 333–341. doi:10.1177/1740774515583488.

Assessing treatment effects with surrogate survival outcomes using an internal validation subsample

Jarcy Zee[1], **Sharon X Xie**[1], and **for the Alzheimer's Disease Neuroimaging Initiative**[2]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA

[2]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/ how to apply/ADNI Acknowledgement List.pdf.

Abstract

Background—In studies with surrogate outcomes available for all subjects and true outcomes available for only a subsample, survival analysis methods are needed that incorporate both endpoints in order to assess treatment effects.

Methods—We develop a semiparametric estimated likelihood method for the proportional hazards model with discrete time data and a binary covariate of interest. Our proposed method allows for real-time validation of surrogate outcomes and flexible censoring mechanisms.

Results—Our proposed estimator is consistent and asymptotically normal. Through numerical studies, we showed that our proposed method for estimating a covariate effect is unbiased compared to the naïve estimator that uses only surrogate endpoints and is more efficient with moderate missingness compared to the complete-case estimator that uses only true endpoints. We further demonstrated the advantages of our proposed method in comparison to existing approaches when there is real-time validation. We also illustrated the use of our proposed method by estimating the effect of gender on time to detection of Alzheimer's disease using data from the Alzheimer's Disease Neuroimaging Initiative.

Conclusions—The proposed method is able to account for the uncertainty of surrogate outcomes by using a validation subsample of true outcomes in estimating a binary covariate effect. The proposed estimator can outperform standard semiparametric survival analysis methods, and can therefore save on costs of a trial or improve power in detecting treatment effects.

Keywords

Semiparametric survival analysis; proportional hazards model; surrogate outcomes; validation subsample; measurement error; missing data

Corresponding author: Sharon X Xie, Department of Biostatistics and Epidemiology, University of Pennsylvania, 607 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA sxie@mail.med.upenn.edu Phone: 215-573-3867.

Introduction

In clinical trials, interest often lies in comparing the effects of treatment on time to an event. The Cox proportional hazards model is a common method for analyzing true survival outcome data, but true outcomes are often unavailable due to invasiveness or cost restrictions. Surrogate outcomes are often used as an alternative to true outcomes since they are more widely available, but mismeasured surrogate outcomes may produce mismeasured survival estimates. Sometimes, data on both the mismeasured surrogate outcomes and a subset of true outcomes are available.

For example, we consider estimation of the time to pathological detection of Alzheimer's disease (AD), which can be measured by a cerebral spinal fluid (CSF) assay of amyloid beta (Aβ) protein concentrations. Although the gold standard for pathological diagnosis of AD is autopsy, a diagnosis after death does not provide information about time to AD. Furthermore, abnormality of CSF values is highly correlated with autopsy diagnosis and is a well-accepted measure for pathological diagnosis of AD among living participants in research studies,[1] and thus represents the true outcome without error in the current study. However, the CSF assay requires a lumbar puncture to extract spinal fluid, which is considered too invasive for some patients. Therefore a CSF-based outcome has limited availability. Alternatively, clinical detection of AD based on cognitive tests may be used as a surrogate for the pathological detection of AD since it is easier to obtain. However, the clinical symptoms of AD present differently from the pathological signs and therefore measures the true outcome with error. The clinical assessment is more widely available, so we have surrogate outcomes on all subjects, whereas we only have true pathological assessments for some subjects. Using both the mismeasured surrogate outcome on all subjects and the true outcome on a subsample, called the validation sample, estimates of covariate effects can be improved.

Previous methods for estimating mismeasured survival outcomes assumed known mismeasurement rates of the surrogate, such as sensitivity and specificity of the diagnostic test used.[2–5] These previous methods did not incorporate a validation subsample of true outcomes. Pepe[6] developed an estimated likelihood method for data with surrogate outcomes on all subjects and true outcomes on only a subsample, but the method was not specifically for a survival outcome. Magaret[7] extended the estimated likelihood method for discrete survival data without real-time validation using a proportional hazards model. Because validation cannot be conducted in real time, Magaret's method relies on the assumption that true outcomes are censored after false positives and that true and surrogate censoring times are equal when the surrogate outcome is censored (second full paragraph on page 5459 of Magaret[7]). In some situations, however, real-time validation is possible and the true and surrogate outcomes follow separate trajectories. In the AD example, subjects in the validation subsample can undergo regular clinical screenings while they also separately undergo independent CSF testing, so true outcomes are not necessarily censored after false positives. Therefore, the time to pathological AD detection may be before or after the time to clinical detection. Also, true and surrogate censoring times can be completely different. Zee and Xie[8] adapted the estimated likelihood method[6] to nonparametrically estimate a survival function assuming real-time validation is possible. The method does not assume

known mismeasurement rates of the surrogate outcome and allows for flexible censoring mechanisms.

In this paper, we extend the work of Zee and Xie[8] to a semiparametric estimated likelihood method to estimate a parameter representing a binary covariate effect for discrete survival data with surrogate outcomes on all subjects and true outcomes on a subsample. Although we express our approach with a binary variable for ease of notation, the method can be easily modified for categorical variables with more than two levels. The rest of the article is organized as follows. We first describe the estimated likelihood and asymptotic properties for the estimated effect of a binary covariate. The Simulation study section contains results from testing the performance of our proposed estimator. In the Data example section, we demonstrate the use of our method using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to estimate the effect of gender on time to detection of pathological AD. Finally, we summarize our results and discuss implications of using our proposed method in the Discussion.

Semiparametric estimated likelihood with a binary covariate

Maximum estimated likelihood estimation—We let T represent the true time to event and C represent the true right censoring time. The true observed time is represented by $X = \min\{T, C\}$ and true observed event indicator by $\delta = I(T \leq C)$. Similarly, the surrogate outcome is denoted with asterisks, with T^* and C^* representing the surrogate event and censoring times, respectively, $X^* = \min\{T^*, C^*\}$ representing the surrogate observed time, and $\delta^* = I(T^* \leq C^*)$ representing the surrogate event indicator. We let X_k represent the kth unique, ordered observed true time point for $k = 1, \ldots, K$, where K is the total number of unique true observed times. Let F_0 represent the baseline survival function of the true time to event. We assume a proportional hazards model with $F(t) = F_0(t)^{\exp(\beta Z)}$ where $Z \in \{0, 1\}$ is the binary covariate of interest and represents the log hazard ratio comparing $Z = 1$ to $Z = 0$. We assume that the covariate is available for all subjects. Finally, to allow for random censoring, we let G represent the censoring survival function.

As in the standard Cox model, we assume independent censoring conditional on covariates, and we allow the censoring mechanism to be fixed or random. Fixed, or type 1, censoring refers to a special case of administrative censoring where all subjects enter and leave the study at the same time, so censoring time is known at the start of the study and is equal for everyone. Random censoring refers to the situation where censoring time is unknown at start of study and may occur randomly due to circumstances such as loss to follow-up. A third type of censoring, sometimes known as generalized type 1 censoring, occurs when censoring time is known in advance but subjects enter the study at random times. In this case, subjects who are censored have different, random observation times and are therefore also called randomly censored in this paper. Because we assume real-time validation is possible, the true and surrogate outcomes can have different censoring times, which may occur if a subject drops out of only one part of a study. In the AD example, a subject may agree to both clinical and CSF screenings at the start of study, but after some time opt out of only the CSF screenings. Here, the subject's true censoring time may occur before the surrogate censoring time.

Let V represent the validation set, the set of subjects for whom both the surrogate and true outcomes are available. Let $\bar{V}$ represent the non-validation set, in which only the surrogate outcome is available and the true outcome is missing. There are n total subjects in the sample and n_V in V. We assume V is a representative sample of the entire study cohort, i.e., those missing the true outcome are missing completely at random (MCAR). The estimated likelihood is a function of the log hazard ratio, β, and possible survival function values for the baseline event distribution and censoring distribution at each time point. In semiparametric survival analysis, the parameter of interest is often only the log hazard ratio, so the survival functions can be considered nuisance parameters. Using similar arguments as in Pepe and Zee and Xie,[6;8] the estimated likelihood is given by

$$\hat{L}(\beta, F_0, G) \propto \prod_{i \in V} P(X_i, \delta_i | Z_i) \prod_{j \in \bar{V}} \hat{P}\left(X_j^*, \delta_j^* | Z_j\right).$$

where

$$\hat{P}\left(X_j^*, \delta_j^* | Z_j\right) = \sum_{k=1}^{K} \sum_{\delta=0}^{1} P(x_k, \delta | Z_j)\, \hat{P}\left(X_j^*, \delta_j^* | x_k, \delta, Z_j\right).$$

The outer sum is summed over all possible time points and the inner sum is summed over all possible event indicators. The conditional probability is estimated empirically with proportions,

$$\begin{aligned} \hat{P}\left(X_j^*, \delta_j^* | x_k, \delta, Z_j\right) &= \frac{\hat{P}\left(X_j^*, \delta_j^*, x_k, \delta, Z_j\right)}{\hat{P}(x_k, \delta, Z_j)} \\ &= \frac{\frac{1}{n_V} \sum_{i \in V} I\left(X_j^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta, Z_i = Z_j\right)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta, Z_i = Z_j)} \end{aligned}$$

where $I(\cdot)$ is the indicator function. Given surrogate outcome and covariate values of a non-validation set subject and given each possible true outcome value, the estimated conditional probability is calculated by counting the number of subjects in the validation set with that set of values, out of the number of subjects in the validation set with that set of true outcome and covariate values. Therefore, the values observed in the validation set are used to estimate the association between the true and surrogate outcomes to determine the likelihood contributions of the non-validation set subjects.

The form of the empirical probabilities assumes that the covariate values are important in estimating the relationship between the true and surrogate outcomes. However, in cases where $P\left(X_j^*.\delta_j^* | x_k, \delta, Z_j\right) = P\left(X_j^*, \delta_j^* | x_k, \delta\right)$, or when the covariate is uninformative about the association between outcomes, then the covariates can be removed from the indicator functions in the probability estimates.

For subjects $i \in V$, the contribution to the likelihood is

$$P\left(X_i,\delta_i|Z_i\right) = \left\{F_0(x_{k_i-1})^{exp(\beta Z_i)} - F_0(x_{k_i})^{exp(\beta Z_i)}\right\}^{\delta_i}\left\{F_0(x_{k_i})^{exp(\beta Z_i)}\right\}^{1-\delta_i} \cdot G(x_{k_i-1})^{\delta_i}\{G\left(x_{k_i-1}\right) - G\left(x_{k_i}\right)\}^{1-\delta_i} \propto \left\{F_0(x_{k_i-1})^{exp(\beta Z_i)} - F_0(x_{k_i})^{exp(\beta Z_i)}\right\}^{\delta_i}\left\{F_0(x_{k_i})^{exp(\beta Z_i)}\right\}^{1-\delta_i}$$

where x_{ki} is the observed time for subject i. This expression is exactly what it would be for a standard proportional hazards model. For subjects $j \in V$, the contribution to the likelihood is

$$\hat{P}\left(X_j^*,\delta_j^*|Z_j\right) = \sum_{k=1}^{K}\sum_{\delta=0}^{1}\left[\left\{F_0(x_{k-1})^{exp(\beta Z_j)} - F_0(x_k)^{exp(\beta Z_j)}\right\}^{\delta}\left\{F_0(x_k)^{exp(\beta Z_j)}\right\}^{1-\delta} \cdot G(x_{k-1})^{\delta}\{G\left(x_{k-1}\right) - G\left(x_k\right)\}^{1-\delta} \cdot \frac{\frac{1}{n_V}\sum_{i\in V} I\left(X_i^*=X_j^*,\delta_i^*=\delta_j^*,X_i=x_k,\delta_i=\delta,Z_i=Z_j\right)}{\frac{1}{n_V}\sum_{i\in V} I(X_i=x_k,\delta_i=\delta,Z_i=Z_j)}\right].$$

The marginal distribution of the true outcome given covariates, $P(x_k, \delta|Z_j)$, for the non-validation set has the same form as it did for the validation set. However, unlike in the validation set contribution, the outer sum prevents the censoring distribution from being factored out, which is necessary to obtain a consistent estimator for β in the presence of random censoring.

Although the parameter of interest is often only β, we maximize the estimated likelihood jointly over all possible parameter values. As in the nonparametric case, the maximum estimated likelihood estimate for the event (censoring) survival function is a step function that falls only at event (censoring) times observed in the validation set. We solve for maximum estimated likelihood estimates using the Nelder-Mead algorithm with constraints on both survival functions to be monotonically non-increasing and bounded between 0 and 1. To obtain initial estimates for the event distribution parameters, we used the complete-case Kaplan-Meier estimates based on the true observed times and true event indicators from the validation set. Initial parameters for the censoring distribution were determined by the complete-case Kaplan-Meier estimates calculated by inverting the event indicator to obtain a censoring indicator. The initial parameter for the covariate effect is set at 0 and is unconstrained.

Asymptotic properties of $\hat{\beta}$—We assume that the proportion of subjects in the validation set out of the total number of subjects does not have a zero limit, or $lim_{n\to\infty}\frac{n_v}{n}=p_V>0$. Then, similar arguments as in Theorem 3.1 of Pepe[6] imply that $\hat{\beta}$ is a consistent estimator for β as $n \to \infty$ and

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{D} N\left(0,\sigma^2\right)$$

where σ^2 is the [1,1] element of the full variance covariance matrix

$$\Sigma = \mathscr{I}^{-1} + \frac{(1-p_V)^2}{p_V}\mathscr{I}^{-1}\mathscr{K}\mathscr{I}^{-1},$$

where $\mathscr{I}$ is the information matrix based on the (non-estimated) log likelihood and $\mathscr{K}$ is the expected conditional variance of the score function of the non-validation contribution to the log likelihood,[6]

$$\mathscr{K} = \mathrm{E}\left[Var\left\{\left.\frac{\partial \log P(X^*, \delta^* | Z)}{\partial\theta}\right| X, \delta, Z\right\}\right]$$

for parameters $\theta = \{\beta, F_0, G\}$. The first term in the Σ matrix represents the variance based on the maximum likelihood estimator, as in standard maximum likelihood analysis. The second term of Σ is needed in order to account for the additional variability introduced by estimating the likelihood with empirical probabilities. The $\mathscr{I}$ and $\mathscr{K}$ matrices can be estimated consistently by the expressions

$$\hat{\mathscr{I}}\frac{1}{n}\left.\frac{\partial^2 \log \hat{L}}{\partial\theta^2}\right|_{\theta=\hat{\theta}}$$

for maximum estimated likelihood estimates $\hat{\theta} = \left\{\hat{\beta}, \hat{F}_0, \hat{G}\right\}$ and

$$\hat{\mathscr{K}} = \frac{1}{n_V}\sum_{i\in V}\left.\hat{Q}_i, \hat{Q}_i^T\right|_{\theta=\hat{\theta}},$$

where

$$\hat{Q}_i = \frac{1}{n-n_V}\frac{1}{\hat{P}(X_i,\delta_i,Z_i)}\sum_{j\in\bar{V}}\left[\left\{I\left(X_j^*=X_i^*, \delta_j^*=\delta_i^*\right) - \hat{P}\left(X_j^*, \delta_j^* | X_i, \delta_i, Z_i\right)\right\} I(Z_i=Z_j) \cdot \left\{\frac{D(X_i,\delta_i|Z_j)}{\hat{P}\left(X_j^*,\delta_j^*|Z_j\right)} - \frac{\hat{D}\left(X_j^*,\delta_j^*|Z_j\right)}{\hat{P}^2\left(X_j^*,\delta_j^*|Z_j\right)}P(X_i,\delta_i|Z_j)\right\}\right]$$

and

$$\begin{aligned}
\hat{P}(X_i,\delta_i,Z_i) &= \frac{1}{n_V}\sum_{a\in V} I(X_a=X_i, \delta_a=\delta_i, Z_a=Z_i)\\
\hat{P}\left(X_j^*,\delta_j^*|X_i,\delta_i,Z_i\right) &= \frac{\frac{1}{n_V}\sum_{a\in V} I\left(X_a^*=X_j^*,\delta_a^*=\delta_j^*,X_a=X_i,\delta_a=\delta_i,Z_a=Z_i\right)}{\frac{1}{n_V}\sum_{a\in V} I(X_a=X_i,\delta_a=\delta_i,Z_a=Z_i)}\\
D(X_i,\delta_i|Z_j) &= \frac{\partial P(X_i,\delta_i|Z_j)}{\partial\theta}\\
\hat{D}\left(X_j^*,\delta_j^*|Z_j\right) &= \sum_{k=1}^{K}\sum_{\delta=0}^{1}\frac{\partial P(x_k,\delta|Z_j)}{\partial\theta}\hat{P}\left(X_j^*,\delta_j^*|x_k,\delta,Z_j\right).
\end{aligned}$$

In practice, we can calculate the estimated variance with numerical derivatives, analytical forms for derivatives, or using bootstrapping. As in the nonparametric case, we found that the numerical derivatives were sometimes incalculable with large amounts of missing data or a large number of parameters to estimate.

Simulation study

To test the performance of our proposed method in estimating a covariate effect, we conducted a series of simulations. We sampled values Z~Bernoulli(0.5) for the binary covariate and set the log hazard ratio at $\beta = 1$. We sampled true event times assuming a proportional hazards model with baseline distribution, $T|Z = 0$ ~Unif[1, 5], where survival time can only take integer values. For now, we assumed fixed right censoring at $C = 4$. The surrogate time to event was calculated as $T^* = T + \varepsilon$, where ε ~Unif[0,ζ] and ε is independent of T . The maximum integer value of the discrete uniform distribution for ε was calculated as $\zeta = \left\lfloor \sqrt{Var(T) \cdot \frac{1-\rho^2}{\rho^2} + 1 - 1} \right\rfloor$, where $\lfloor a \rfloor$ represents the largest integer not greater than a and ρ represents the correlation between T and T^* . We considered correlations of $\rho \in \{0.01, 0.25, 0.50, 0.75, 1\}$. We set the right-censoring time for the surrogate endpoint to be fixed also at $C^* = 4$. To create a representative validation subsample, we simulated data MCAR by randomly selecting a proportion $r \in \{0.25, 0.50\}$ of the sample to be missing true outcomes.

We conducted 500 simulation repetitions for each set of parameter values and used a total sample size of $n = 500$ for each repetition. For each simulated dataset, we used the proposed method to calculate estimates of the log hazard ratio, $\hat{\beta}$. We also calculated the complete-case estimate of the log hazard ratio by using only available true outcomes in the validation set, the naïve estimate using only surrogate outcomes from all subjects, and the true estimate using true outcomes from all subjects (which would not be possible with real data but is consistent and optimally efficient). For each of the three standard estimators, we used the maximum likelihood estimates (MLEs) rather than partial likelihood estimates. Although partial likelihood estimates are more common in practice, the MLEs are more accurate and compare better to our proposed method which is also an MLE method. For each method, we calculated estimated bias (parameter estimate−true parameter values), observed sample standard deviations (SD), estimated standard errors $\left(\hat{SE}\right)$, relative e ciency (RE) compared to the true estimator (where lower RE implies greater efficiency and RE equal to 1 implies optimal efficiency), mean squared error (MSE) estimates, and 95% coverage (Cov). We note that for all simulations presented in Tables 1 and 2, the observed sample standard deviations are similar to the standard error estimates calculated using the asymptotic theory for the proposed estimator.

Table 1 shows the results from the simulations with fixed (type 1) censoring. The log hazard ratio estimates estimated by our proposed method and the complete-case estimator always have little bias, whereas the naïve estimates are biased whenever the correlation between outcomes is less than 1. In a few cases, the standard errors for our proposed estimator were slightly higher than the complete-case estimator due to the penalty that is added to the variance of our proposed estimator for estimating the likelihood. However, the penalty is

small and overall our proposed estimator has similar standard errors compared to the complete-case estimator when the correlation between outcomes is low. When the correlation between outcomes increases, though, our proposed estimator is able to incorporate more information from the non-validation set subjects and therefore improves in efficiency, far surpassing the added penalty. At correlation of 1, which can be interpreted as the situation where we have a perfect surrogate, our proposed estimator has optimal efficiency.

As we observed in the nonparametric case, the efficiency gains of our proposed estimator changed with different amounts of missingness and correlation between true and surrogate outcomes. We conducted additional simulations to explore and confirm this relationship by testing amounts of missingness between 0 and 80. We found that our proposed estimator was more efficient than the complete-case estimator when the missingness was low. With high missingness, our proposed estimator became less efficient than the complete-case estimator. The point at which our estimator crosses from more to less efficient differs with the correlation between outcomes—as correlation between true and surrogate outcomes increases, the amount of missingness at the crossing point increases. Sometimes the correlation between outcomes is unknown or unable to be estimated. Regardless of the correlation between outcomes, our simulations showed that our proposed estimator has similar or greater efficiency than the complete-case estimator when the amount of missingness is 50% or less of the total sample. This was consistent with previous work.[8]

The previous simulations assumed measurement error was uniformly distributed and positive. We also conducted simulations assuming measurement error was geometrically distributed or uniformly distributed but could be positive or negative. Results are not shown, but were similar to those seen above.

Although we assumed true outcomes were MCAR, we also conducted additional simulations with data missing at random (MAR) to test the robustness of this assumption. To do so, we assumed that the probability of validation was $p_R \in \{0.60, 0.70, 0.85\}$ for subjects who had a positive surrogate outcome, and the probability of validation for subjects who had a negative surrogate outcome was $1 - p_R$. Under this scenario, we found that our proposed method and the complete-case method were both somewhat biased, particularly with higher values of p_R. However, our proposed estimator was less biased and therefore had better coverage than the complete-case estimator with higher correlations between outcomes.

We also simulated data assuming random censoring and changed the amount of censoring by sampling true censoring times C from a uniform distribution. We considered a small amount of censoring (approximately 17%) using $C \sim \text{Unif}[3, 4]$, a moderate amount of censoring (approximately 36%) using $C \sim \text{Unif}[1, 4]$, and a large amount of censoring (approximately 84%) using $C \sim \text{Unif}[1, 2]$. Surrogate censoring times were simulated by $C^* = C + \gamma$ where $\gamma \sim \text{Unif}[0, 2]$. The results of these random censoring simulations are shown in Table 2. Similar to the results with fixed censoring, our proposed estimator and the complete-case estimator have little bias compared to the naïve estimator and our proposed estimator is more efficient than the complete-case estimator at any amount of censoring.

To demonstrate the utility of our proposed method with real-time validation, we compared our proposed method to Magaret's.[7] We sampled event times assuming a baseline hazard rate of 2, a binary covariate, Z ~Bernoulli(0.5), and assumed a proportional hazards model with log hazard ratio of $\beta = 0.70$. Surrogate event times were calculated as $T^* = T + \varepsilon$, ε ~Unif[0, 2]. We simulated fixed but unequal true and surrogate censoring times, $C = 4$ and $C^* = 5$, and also simulated random censoring by sampling C from Unif[4, 5], Unif[2, 5], and Unif[1, 3], resulting in approximately 25, 35, and 55 percent censoring, respectively. For random censoring, surrogate censoring times were calculated as $C^* = C + \gamma$, $\gamma \sim$ Unif[0, 2]. We used total sample sizes of $n \in \{240, 420, 630\}$ with 50% missingness. For each simulated dataset, we used our proposed method and Magaret's method to calculate estimates of the log hazard ratio. Bias and observed sample variances for these comparisons are shown in Table 3. Our proposed estimator has little bias; however, the method developed by Magaret[7] is biased when we have real-time validation.

Data example: effect of gender on time to pathological detection of Alzheimer's disease

To illustrate our proposed method, we considered data (retrieved on 26 July 2013) from the ADNI study.[9] Participants in this observational study were assessed at predetermined time points for genetic, biomarker, and imaging markers related to AD. A description of the ADNI study is in Appendix 1. In the current study, participants who were non-AD at baseline were included ($n = 186$). Participants had to be non-AD by both a clinical and a CSF-based assessment at baseline to ensure all participants were event free for both outcomes at study entry. For the CSF-based outcome, $A\beta >$ 192pg/ml was classified as non-AD and $A\beta \leq$ 192pg/ml was classified as AD.[1] CSF assessments were completely independent from clinical assessments.

The true outcome of interest was time to pathological detection of AD, measured in years. Since AD is a chronic disease with slow progression[10] and annual follow-up times were predetermined by study design, survival time was considered to be discrete. For every patient, the surrogate outcome of time to clinical AD or last follow-up was determined. A subset of 110 patients continued to have CSF assays performed after baseline, independently from clinical screenings, from which the true time to pathological AD or last follow-up was determined. Therefore, the validation set was approximately 59% of the total sample size. All patients in the study also had information on gender.

Using our proposed method, we estimated the log hazard ratio, $\hat{\beta}$, of AD in females compared to males. We also estimated the log hazard ratio with the complete-case estimator using only 110 CSF diagnoses and the naïve estimator using only 186 clinical diagnoses. For the standard estimators, we conducted estimation using both the maximum likelihood method and the more widely used partial likelihood method with Efron's approximation for ties.[11] Table 4 shows the log hazard ratio and standard error estimates for gender. Both our proposed estimator and the complete-case estimator found a small positive log hazard ratio comparing females to males, which is similar to some literature indicating higher incidence of AD in women.[12;13] However, the naïve estimate is large and negative. In this particular example comparing genders, the estimated standard errors from our proposed method and the complete-case method were similar.

Discussion

We extended the nonparametric estimated likelihood method for data with surrogate endpoints and an internal validation subsample to the proportional hazards model with a binary covariate. Our method allows for real-time validation and allows for flexible censoring mechanisms. Our proposed log hazard ratio estimator is consistent and asymptotically normal. Through simulation studies, we found that our proposed estimate is unbiased and its variance decreases as correlation between the surrogate and true outcome increases. By using both surrogate and true endpoints, the proposed covariate effect estimator can outperform both complete-case and naïve estimators.

As in the nonparametric case, we found that our proposed estimator behaves similarly to the complete-case estimator when the correlation between the uncertain and true outcomes is low. As correlation increases, the non-validation set subjects contribute more information and therefore decrease the variance of the parameter estimate by providing more power. When correlation between outcomes is 1, or when the surrogate outcome has no measurement error, then our proposed estimator reduces to the maximum likelihood estimate based on complete true outcomes (no missingness). Therefore, our proposed method is most useful when correlation between outcomes is high.

In our current study, we evaluated the use of an estimated likelihood method for survival data with a single binary covariate. The method can easily be extended to consider multiple covariates, which would be useful in order to adjust for confounding variables or to consider categorical variables with more than two levels. Further study on the number of allowable covariates is warranted; however, based on the events per variable (EPV) testing in Zee and Xie,[8] it is expected that a similar EPV of 4 would apply to multivariable models. This would imply that a minimum of 4 events should be observed for each parameter to be estimated. For continuous covariates, however, a modified approach must be developed, since likelihood contributions would often be 0 for non-validation set subjects when using the current method. An estimated likelihood method that incorporates smooth kernel functions in the empirical probability estimates of the likelihood function is currently under investigation. Additionally, other applications may involve interval censoring, for which our methods would need to be modified to be suitable.

Optimal study design strategies are currently under investigation, in order to determine the total size of the sample and size of the validation subset that would be needed to design new trials with these data characteristics. For example, in clinical trials that aim to evaluate a treatment effect, it is valuable to calculate sample sizes that would be needed in each treatment group to achieve a pre-specified power to detect the effect size. The improvements in efficiency that can be obtained from using the proposed method would decrease the number of true outcomes needed compared to using standard methods.

Due to the difficulty in obtaining true outcomes on many subjects, the methods we have proposed have useful applications in clinical trials. Designing studies such that surrogate outcomes are collected on all patients and true outcomes only collected on a subsample of patients can save on trial costs and ensure that an adequate number of patients are enrolled.

Using our proposed semiparametric estimated likelihood method to analyze these data can provide accurate and powerful statistical inference to evaluate treatment effects.

Funding Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). Dr. Zee received support from NIH National Institute of Mental Health grant T32MH065218 and Dr. Xie from NIH National Institute on Aging grant AG10124 (University of Pennsylvania Alzheimer's Disease Core Center), AG32953, AG17586, and NS053488.

Appendix 1 ADNI Study Description

The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of ADNI is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of e orts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Ho mann-La Roche Ltd and its affliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Rev December 5, 2013 Health Research is providing funds to support ADNI clinical sites in Canada. Private sector

contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

1. Shaw L, Vanderstichele H, Knapik-Czajka M, et al. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Annals of Neurology. 2009; 65(4):403–413. DOI:10.1002/ana.21610.Cerebrospinal. PMCID: PMC2696350. [PubMed: 19296504]
2. Snapinn SM. Survival analysis with uncertain endpoints. Biometrics. 1998; 54(1):209–18. PMID: 9574966. [PubMed: 9574966]
3. Richardson BA, Hughes JP. Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. Biostatistics. 2000; 1(3):341–54. DOI:10.1093/biostatistics/1.3.341. PMID: 12933514. [PubMed: 12933514]
4. Meier AS, Richardson BA, Hughes JP. Discrete Proportional Hazards Models for Mismeasured Outcomes. Biometrics. 2003; 59(4):947–954. PMCID: PMC2574985. [PubMed: 14969473]
5. Balasubramanian R, Lagakos SW. Estimation of the timing of perinatal transmission of HIV. Biometrics. 2001; 57(4):1048–58. PMID: 11764243. [PubMed: 11764243]
6. Pepe M. Inference using surrogate outcome data and a validation sample. Biometrika. 1992; 79(2):355–365.
7. Magaret AS. Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. Statistics in Medicine. 2008; 27:5456–5470. DOI:10.1002/sim. PMCID: PMC2574985. [PubMed: 18613225]
8. Zee J, Xie SX. Nonparametric Discrete Survival Function Estimation with Uncertain Endpoints Using an Internal Validation Subsample. Biometrics. 2015 in press.
9. Weiner MW, Veitch DP, Aisen PS, et al. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimer's & Dementia. 2012; 8(1 Suppl):S1–68. DOI:10.1016/j.jalz.2011.09.172. PMID: 22047634, PMCID: PMC3329969.
10. Jack CR Jr, Knopman DS, Jagust WJ, et al. Hypothetical Model of Dynamic Biomarkers of the Alzheimer's Pathological Cascade. Lancet Neurology. 2010; 9(1):119–128. DOI:10.1016/S1474-4422(09)70299-6.Hypothetical. [PubMed: 20083042]
11. Efron B. The Efficiency of Cox's Likelihood Function for Censored Data. Journal of the American Statistical Association1. 1977; 72(359):557–565.
12. Mielke MM, Prashanthi V, Rocca WA. Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences. Clinical Epidemiology. 2014; 6:37–48. PMCID: PMC3891487. [PubMed: 24470773]
13. Letenneur L, Gilleron V, Commenges D, et al. Are sex and educational level independent predictors of dementia and Alzheimer's disease? Incidence data from the PAQUID project. Journal of Neurology, Neurosurgery, & Psychiatry. 1999; 66:177–183. PMCID: PMC1736218.

Table 1

Simulation Results for Fixed Censoring

r	*ρ*	Method	Bias	SD	$\widehat{SE}$	MSE	RE	Cov
25	0.01	Proposed	0.013	0.124	0.119	0.016	1.44	0.94
		Comp	0.010	0.123	0.119	0.015	1.42	0.95
		Naïve	−1.115	0.697	0.912	1.731	45.47	0.92
	0.25	Proposed	0.013	0.124	0.119	0.016	1.44	0.94
		Comp	0.010	0.123	0.119	0.015	1.42	0.95
		Naïve	−0.510	0.251	0.246	0.323	5.89	0.41
	0.50	Proposed	0.012	0.122	0.117	0.015	1.39	0.94
		Comp	0.010	0.123	0.119	0.015	1.42	0.95
		Naïve	−0.445	0.170	0.163	0.227	2.70	0.23
	0.75	Proposed	0.008	0.115	0.113	0.013	1.24	0.95
		Comp	0.010	0.123	0.119	0.015	1.42	0.95
		Naïve	−0.223	0.124	0.116	0.065	1.44	0.51
	1.00	Proposed	0.007	0.106	0.103	0.011	1.04	0.94
		Comp	0.010	0.123	0.119	0.015	1.42	0.95
		Naïve	0.004	0.103	0.103	0.011	1.00	0.94
50	0.01	Proposed	0.038	0.158	0.146	0.026	2.34	0.93
		Comp	0.026	0.153	0.147	0.024	2.18	0.95
		Naïve	−1.105	0.715	0.917	1.733	47.81	0.92
	0.25	Proposed	0.040	0.160	0.146	0.027	2.40	0.92
		Comp	0.026	0.153	0.147	0.024	2.18	0.95
		Naïve	−0.510	0.251	0.246	0.323	5.89	0.41
	0.50	Proposed	0.033	0.154	0.143	0.025	2.21	0.93
		Comp	0.026	0.153	0.147	0.024	2.18	0.95
		Naïve	−0.445	0.170	0.163	0.227	2.70	0.23
	0.75	Proposed	0.020	0.137	0.130	0.019	1.76	0.94
		Comp	0.026	0.153	0.147	0.024	2.18	0.95
		Naïve	−0.223	0.124	0.116	0.065	1.44	0.51
	1.00	Proposed	0.016	0.110	0.103	0.012	1.13	0.93
		Comp	0.026	0.153	0.147	0.024	2.18	0.95
		Naïve	0.004	0.103	0.103	0.011	1.00	0.94

Note: True β=1; *r*: percent missing; ρ: correlation between true and surrogate outcomes; Proposed: proposed estimator; Comp: complete-case estimator; Naïve: naïve estimator; SD: standard deviation of estimates across simulations; $\widehat{SE}$: estimated standard error of the estimate; MSE: mean squared error; RE: relative efficiency compared to true estimator; Cov: 95% coverage.

Table 2

Simulation Results for Random Censoring and a Binary Covariate

r	*C*	Method	Bias	SD	$\widehat{SE}$	MSE	RE	Cov
25	17	Proposed	0.007	0.114	0.112	0.013	1.17	0.94
		Comp	−0.000	0.121	0.122	0.015	1.32	0.95
		Naïve	−0.145	0.110	0.107	0.033	1.09	0.73
	36	Proposed	0.010	0.129	0.129	0.017	1.14	0.94
		Comp	0.009	0.135	0.140	0.018	1.25	0.96
		Naïve	−0.125	0.132	0.123	0.033	1.19	0.80
	84	Proposed	0.005	0.161	0.157	0.026	1.18	0.95
		Comp	0.004	0.165	0.166	0.027	1.25	0.95
		Naïve	−0.132	0.149	0.153	0.040	1.01	0.86
50	17	Proposed	0.023	0.129	0.124	0.017	1.50	0.93
		Comp	0.004	0.152	0.151	0.023	2.08	0.94
		Naïve	−0.145	0.110	0.107	0.033	1.09	0.73
	36	Proposed	0.026	0.151	0.147	0.023	1.56	0.96
		Comp	0.022	0.169	0.173	0.029	1.95	0.96
		Naïve	−0.125	0.132	0.123	0.033	1.19	0.80
	84	Proposed	0.013	0.186	0.184	0.035	1.59	0.95
		Comp	0.003	0.201	0.205	0.040	1.85	0.95
		Naïve	−0.132	0.149	0.153	0.040	1.01	0.86

Note: True β=1; *r*: percent missing; *C*: percent censoring; Proposed: proposed estimator; Comp: complete-case estimator; Naïve: naïve estimator; SD: standard deviation of estimates across simulations; $\widehat{SE}$: estimated standard error of the estimate; MSE: mean squared error; RE: relative efficiency compared to true estimator; Cov: 95% coverage.

Table 3

Simulation Results Compared to Magaret[7]

N	*C*	Method	Bias	Var
240	fixed	Proposed	−0.021	0.031
		Magaret	−0.246	0.026
	25	Proposed	0.010	0.031
		Magaret	−0.123	0.027
	35	Proposed	0.002	0.040
		Magaret	−0.078	0.038
	55	Proposed	−0.008	0.058
		Magaret	−0.158	0.047
420	fixed	Proposed	−0.004	0.018
		Magaret	−0.249	0.013
	25	Proposed	−0.000	0.018
		Magaret	−0.122	0.015
	35	Proposed	0.015	0.022
		Magaret	−0.055	0.022
	55	Proposed	0.013	0.035
		Magaret	−0.140	0.029
630	fixed	Proposed	−0.013	0.011
		Magaret	−0.248	0.009
	25	Proposed	−0.008	0.012
		Magaret	−0.115	0.011
	35	Proposed	−0.001	0.013
		Magaret	−0.066	0.013
	55	Proposed	−0.005	0.019
		Magaret	−0.153	0.018

Note: *C*: type or percent of censoring if random; Proposed: proposed estimator; Magaret: estimator from Magaret;[7] Var: variance of estimates across simulations.

Table 4

Data Example Log Hazard Ratio and Standard Error Estimates

	Proposed	Comp (MLE)	Comp (Partial)	Naïve (MLE)	Naïve (Partial)
Female					
β	0.306	0.164	0.218	−1.609	−1.689
SE	0.630	0.610	0.556	0.839	0.775

Proposed: proposed estimator; Comp: complete-case estimator; Naïve: naïve estimator; MLE: estimated using the maximum likelihood method; Partial: estimated using the partial likelihood method; SE: estimated standard error.