 DOI: 10.1111/biom.12316

Nonparametric Discrete Survival Function Estimation with Uncertain Endpoints Using an Internal Validation Subsample

Jarcy Zee and Sharon X. Xie*

Department of Biostatistics and Epidemiology, University of Pennsylvania, 607 Blockley Hall, 423 Guardian Drive, Philadelphia, Pennsylvania 19104, U.S.A.
**email:* sxie@mail.med.upenn.edu

Summary. When a true survival endpoint cannot be assessed for some subjects, an alternative endpoint that measures the true endpoint with error may be collected, which often occurs when obtaining the true endpoint is too invasive or costly. We develop an estimated likelihood function for the situation where we have both uncertain endpoints for all participants and true endpoints for only a subset of participants. We propose a nonparametric maximum estimated likelihood estimator of the discrete survival function of time to the true endpoint. We show that the proposed estimator is consistent and asymptotically normal. We demonstrate through extensive simulations that the proposed estimator has little bias compared to the naïve Kaplan–Meier survival function estimator, which uses only uncertain endpoints, and more efficient with moderate missingness compared to the complete-case Kaplan–Meier survival function estimator, which uses only available true endpoints. Finally, we apply the proposed method to a data set for estimating the risk of detecting Alzheimer's disease from the Alzheimer's Disease Neuroimaging Initiative.

Key words: Measurement error; Missing data; Nonparametric survival analysis; Uncertain endpoints; Validation sample.

1. Introduction

Survival function estimation is crucial in studying disease progression and therapeutic benefits of drugs in epidemiology studies and clinical trials that involve time-to-event data. However, event outcomes may be subject to measurement error, which can lead to misclassification of the true event outcome. Gold standard or better outcome measurements are sometimes unavailable due to high costs or invasive procedures, and using only complete, true outcomes may exclude many subjects due to missing data. For example, the pathological diagnosis of Alzheimer's disease (AD) has been traditionally determined by autopsy. Recently, as we enter the exciting new era of "personalized medicine," AD biomarker research has been very successful. It is well accepted now that time to pathological diagnosis of AD can be reliably measured by time to an abnormal biomarker value among living participants in research studies (Shaw et al., 2009). Specifically, the amyloid beta ($A\beta$) protein biomarker from a cerebral spinal fluid (CSF) assay has been shown to represent the pathological aspects of AD well and the abnormality of $A\beta$ can be used as a reliable (true) endpoint for studying time to pathological diagnosis of AD among living participants (Shaw et al., 2009). However, the CSF biomarker assay involves a lumbar puncture, so it is often considered too invasive for many patients and therefore has limited availability. An alternative outcome is time to diagnosis of AD by clinical assessment, which relies primarily on cognitive tests. The clinical diagnosis is widely available, but it measures the outcome of pathological diagnosis with error. Sources of error in clinical diagnosis include normal aging independent of AD, "cognitive reserve" due to education-linked factors, and disease heterogeneity (Nelson et al., 2012). Thus, the clinical diagnosis is an uncertain endpoint. Under these circumstances, it is important to develop powerful analytical approaches to use combined information from both true and uncertain endpoints to obtain consistent and more efficient estimators compared to the naïve estimator, which ignores true endpoint measures, and the complete-case estimator, which uses only the available true endpoint measures.

Our proposed method is motivated by survival function estimation of time to pathological development of AD using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Weiner et al., 2012). Participants in the ongoing ADNI study were evaluated at predetermined time points to assess AD development based on cognitive tests. Regardless of these clinical diagnoses, a subset of participants also had longitudinal CSF assays to measure $A\beta$ values, from which time to CSF diagnoses could be determined. Some study participants randomly withdrew from the study before developing cognitive or pathological signs of AD. Therefore, survival time is a discrete random variable subject to random right censoring. Although several nonparametric and semiparametric methods for estimating survival when the outcome is uncertain have been proposed, many rely on prior knowledge of the mismeasurement rates of the uncertain endpoint without an internal validation subsample of true endpoints (Snapinn, 1998; Richardson and Hughes, 2000; Meier, Richardson, and Hughes, 2003; Balasubramanian and Lagakos, 2001). Among those that do incorporate a validation subsample, the method primarily focused on the discrete proportional hazards model in which real-time validation of uncertain outcomes is not possible (Magaret, 2008).

Specifically, Snapinn (1998) estimated weights representing certainty of potential endpoints to modify the Cox proportional hazards model. Richardson and Hughes (2000) obtained unbiased product limit estimates of time to an event with a mismeasured event indicator using an Expectation-Maximization (EM) algorithm. Their estimate uses known information about the sensitivity and specificity of the diagnostic test for having the event without a validation sample. Meier et al. (2003) extended this work for the adjusted proportional hazards model for discrete failure times, also assuming known sensitivity and specificity. Similarly, Balasubramanian and Lagakos (2001) assumed a known time-dependent sensitivity function to estimate the distribution of the time to perinatal HIV transmission.

Pepe (1992) developed an estimated likelihood method to incorporate both uncertain endpoints and a validation subsample to make inference without assuming known sensitivity or specificity. However, the method was not specifically for a survival setting and therefore did not incorporate censoring or the estimation of an entire function over time. Fleming et al. (1994) used Pepe's method for the proportional hazards model by incorporating a validation set available on all subjects (i.e., no missing true endpoint measures) to augment the likelihood for subjects with censored failure times. Magaret (2008) also extended Pepe's work to the discrete proportional hazards model in a method designed for situations where uncertain outcomes could only be validated at the end of study. Therefore, these previous methods are unable to address the unique challenges seen in the ADNI data.

We propose a nonparametric discrete survival function estimator for data with characteristics similar to those of the ADNI study. There are four new contributions to the literature from this article which we summarize below. First, we propose a nonparametric discrete survival function estimator without assuming known mismeasurement rates of the uncertain outcome. Instead, we incorporate both an uncertain outcome on all subjects and a validation subsample of true outcomes to construct a survival estimator using an estimated likelihood method (Pepe, 1992). The standard estimated likelihood method estimates a finite number of parameters, but we use the parameter estimates to define and estimate a survival function over all time by proving that the survival function is a step function. The proposed estimator is the nonparametric maximum estimated likelihood survival function estimator, which is computed by a constrained maximization procedure. In addition, because study subjects are evaluated at predetermined time points by study design, survival time is a discrete random variable for both true and uncertain endpoints. We develop the asymptotic distribution theory of the estimator at all possible time points and provide an asymptotic variance estimator. Second, the proposed nonparametric survival function estimator allows real-time validation and allows missingness or censoring of the true endpoint regardless of the value of the uncertain event indicator. In other words, validation can be conducted at any time during the study and on subjects with either events or censored uncertain outcomes. Third, the proposed estimator is able to handle both type 1 and random right censoring mechanisms for the survival outcomes. Unlike in a standard survival setting, allowing for random censoring involves estimation of a censoring distribution. The censoring distribution parameters are considered nuisance parameters, but they complicate the form of the likelihood and the computation. Fourth, we conducted in-depth exploration of the estimated likelihood estimator's properties, including the effect of correlation between outcomes and amount of missingness on the efficiency of the proposed estimator compared to standard methods, the number of events needed per parameter, and the robustness of missingness assumptions.

We organize the rest of the article as follows. We first describe the estimated likelihood and nonparametric maximum estimated likelihood estimator (Section 2). We then develop the asymptotic properties of the proposed estimator (Section 3). We perform extensive simulations to assess the performance of our proposed estimator and compare it to the complete-case and naïve Kaplan–Meier survival function estimators (Section 4). The simulations consider different correlations between true and uncertain endpoints, amounts and types of censoring, amounts of missingness of true endpoints, types of measurement error, as well as different sample sizes. This is followed by an application to the estimation of the survival function of time to pathological diagnosis of AD using data from the ongoing ADNI study (Section 5). Finally, we summarize our findings and point to applications where incorporating both true and uncertain endpoints are particularly useful (Section 6).

2. Proposed Nonparametric Maximum Estimated Likelihood Estimator

Let T represent the true time to event and C represent the true right censoring time, with event indicator $\delta = I(T \leq C)$. Similarly, let T^* represent the uncertain time to event and C^* be the uncertain right censoring time, with indicator $\delta^* = I(T^* \leq C^*)$. Define $X = \min\{T, C\}$ and $X^* = \min\{T^*, C^*\}$. Then X and X^* represent the true and uncertain observed times, respectively. Let x_k represent the kth ordered true observed time point for $k = 1, \ldots, K$, where K is the total number of discrete true time points that can be observed (i.e., the maximum number of all possible true observed times). Let F represent the survival function of the true time to event and let G represent the survival function of the true censoring time.

Let V represent the validation set, where both the uncertain and true outcomes are available. There are n_V subjects in the validation set. It is assumed that the validation subsample is a representative sample of the entire cohort, implying that data are missing completely at random. Then $\overline{V}$ is the non-validation set, where only the uncertain outcome is available and the true outcome is missing. With a total of n subjects in the study, there are $n - n_V$ subjects in the non-validation set. The entire observed data are $(X_i, \delta_i, X_i^*, \delta_i^*)$ for $i = 1, \ldots, n_V$ and (X_j^*, δ_j^*) for $j = 1, \ldots, n - n_V$. Using similar arguments as in Pepe (1992), the full likelihood would then be

$$L = \prod_{i \in V} P(X_i, \delta_i) P(X_i^*, \delta_i^* | X_i, \delta_i) \prod_{j \in \overline{V}} P(X_j^*, \delta_j^*). \qquad (1)$$

To avoid having to specify or assume the form of the relationship between the true and uncertain endpoints, we propose to

use the estimated likelihood

$$\hat{L} = \prod_{i \in V} P(X_i, \delta_i) \widehat{P}(X_i^*, \delta_i^* | X_i, \delta_i) \prod_{j \in \overline{V}} \widehat{P}(X_j^*, \delta_j^*), \tag{2}$$

where for discrete data,

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

The sum marginalizes the joint distribution to obtain the marginal distribution of the uncertain outcome, so the outer sum is taken over all possible time points, $k = 1, \ldots, K$, and the inner sum over all possible event indicator values. The estimated conditional probability $\widehat{P}(X_j^*, \delta_j^* | x_k, \delta)$ is given by

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

where $I(\cdot)$ is the indicator function. Conceptually, the conditional probability is estimated empirically by counting the proportion of subjects in the validation set whose uncertain outcomes match those of the given non-validation set subject. Because the conditional probability $\widehat{P}(X_i^*, \delta_i^* | X_i, \delta_i)$ from the validation set contribution does not contain any parameters of interest, it can be factored out and the estimated likelihood to be maximized becomes

$$\widehat{L} \propto \prod_{i \in V} P(X_i, \delta_i) \prod_{j \in \overline{V}} \widehat{P}(X_j^*, \delta_j^*). \tag{3}$$

Then for a subject $i \in V$, the contribution to the likelihood is the same as it would be in a standard discrete survival setting,

$$\begin{aligned} P(X_i, \delta_i) &= \{F(x_{k_i - 1}) - F(x_{k_i})\}^{\delta_i} F(x_{k_i})^{1 - \delta_i} G(x_{k_i - 1})^{\delta_i} \{G(x_{k_i - 1}) \\ &\quad - G(x_{k_i})\}^{1 - \delta_i} \\ &\propto \{F(x_{k_i - 1}) - F(x_{k_i})\}^{\delta_i} F(x_{k_i})^{1 - \delta_i} \end{aligned} \tag{4}$$

where x_{k_i} is the observed time for subject i corresponding to the kth time point. Only the true outcome contributes to the likelihood for those in the validation set, implying that uncertain outcomes do not provide any additional information when the true outcome is known. However, the uncertain outcomes for those in the validation set are still used to estimate the relationship between the uncertain and true outcomes, which are then used to weight likelihood contributions for those in the non-validation set. For a subject $j \in \overline{V}$, the contribution to the likelihood is

$$\begin{aligned} &\widehat{P}(X_j^*, \delta_j^*) \\ &= \sum_{k=1}^{K} \sum_{\delta=0}^{1} \left[\{F(x_{k-1}) - F(x_k)\}^{\delta} F(x_k)^{1-\delta} G(x_{k-1})^{\delta} \{G(x_{k-1}) \right. \\ &\quad \left. - G(x_k)\}^{1-\delta} \cdot \frac{\frac{1}{n_V} \sum_{i \in V} I(X_i^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta)} \right]. \end{aligned} \tag{5}$$

Unlike in the validation set contribution, the censoring distribution cannot be factored out of the likelihood from the non-validation set contribution. This distribution is important in allowing random censoring for survival outcomes in the estimated likelihood method. Details on the derivation of the estimated likelihood are available in Web Appendix A. Note that any subjects in the non-validation set with an observed uncertain time that does not match any observed uncertain times in the validation set do not contribute to the likelihood.

The estimated likelihood is a function of possible survival function values for the event distribution and censoring distribution at each time point. The parameters representing the censoring distribution G are estimated jointly with the parameters representing the event distribution F, but treated as nuisance parameters. When the study only has type 1 right censoring, though, the contribution to the likelihood by the censoring distribution will always be 1, so the censoring distribution can be factored out of the likelihood and does not need to be estimated. In order to solve for the nonparametric maximum estimated likelihood survival function estimator F using the estimated likelihood we developed, we first note that the maximum estimate will be a step function that is continuous from the right with left limits and falls only at event times actually seen in the validation set, $t_{\tilde{k}}, \tilde{k} = 1, \ldots, \widetilde{K}$, where $\widetilde{K}$ is the total number of unique true event times actually seen in the data set. Similarly, if the censoring distribution is being estimated, the maximum estimator will be a step function that is continuous from the right with left limits and falls only at censoring times seen in the validation set. Details on the derivation of the step function are available in Web Appendix A. To solve for the parameters, we used the Nelder–Mead algorithm to conduct constrained maximization. We required that both F and G survival functions are monotonically non-increasing as time increases and are bounded between 0 and 1. In the case where the parameter space is one-dimensional, meaning there is only one event time in the validation set data and only type 1 censoring, we used the Brent algorithm. 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. Let $\widehat{F}(t_{\tilde{k}})$ represent the event distribution estimates obtained from the algorithm for $\tilde{k} = 1, \ldots, \widetilde{K}$. The maximum estimated likelihood survival function estimator is then the step function that takes value 1 in the interval $[0, t_1)$,

$\widehat{F}(t_{\tilde{k}})$ in each interval $[t_{\tilde{k}}, t_{\tilde{k}+1})$ for $\tilde{k} = 1, \ldots, \widetilde{K} - 1$, and $\widehat{F}(t_{\tilde{k}})$ in the interval $[t_{\tilde{k}}, x_{\tilde{k}}]$, where $x_{\tilde{k}}$ is the last true observed time and may be equal to $t_{\tilde{k}}$ if a true event occurs at the last true observed time. The estimator is considered undefined after $x_{\tilde{k}}$.

3. Asymptotic Properties of the Proposed Nonparametric Maximum Estimated Likelihood Estimator

The asymptotic properties of the proposed estimator refer to the situation when the total number of subjects $n \to \infty$. As long as the proportion of subjects in the validation set to the total number of subjects does not have a zero limit, $\lim_{n\to\infty} \frac{n_V}{n} = p_V > 0$, similar arguments as in Theorem 3.1 of Pepe (1992) imply that $\widehat{F}(t)$ is a consistent estimator for $F(t)$ for all possible true event times that can be seen. Although $\widehat{F}(t)$ only drops at true event times actually seen in the data, $t_{\tilde{k}}$, $\tilde{k} = 1, \ldots, \widetilde{K}$, this set will approach the set of all possible true event times that can be seen in data, $t_{\mathring{k}}$, $\mathring{k} = 1, \ldots, \mathring{K}$, where $\mathring{K}$ is the total number of possible true event times that can be seen in data. In other words, $\widetilde{K} \to \mathring{K}$ as $n \to \infty$. $\mathring{K}$ is always less than or equal to K and greater than or equal to $\widetilde{K}$. $\mathring{K}$ equals K if all possible true observed times can be events, which is likely to be true in most situations. Note that $\widehat{F}$ is not consistent at all event times if the upper bound of T is greater than the upper bound of C. In this case, event times between the upper bound of C and the upper bound of T can never be seen in data, so $\widehat{F}$ is not consistent at those times and those times are not included in $\mathring{K}$ or in K. Because we have discrete time points, $F(t)$ is also a step function that can be defined by the survival function values at each true event time that can be seen in data, $t_{\mathring{k}}$, $\mathring{k} = 1, \ldots, \mathring{K}$, and we have that

$$\sqrt{n}\left\{\begin{pmatrix}\widehat{F}(t_1)\\ \widehat{F}(t_2)\\ \vdots\\ \widehat{F}(t_{\mathring{K}})\end{pmatrix} - \begin{pmatrix}F(t_1)\\ F(t_2)\\ \vdots\\ F(t_{\mathring{K}})\end{pmatrix}\right\}$$

converges to a zero-mean Gaussian random variable in distribution with asymptotic variance covariance matrix equal to Σ_F, where Σ_F is the top left $\mathring{K} \times \mathring{K}$ quadrant of the full variance covariance matrix

$$\Sigma = \mathcal{I}^{-1} + \frac{(1 - p_V)^2}{p_V}\mathcal{I}^{-1}\mathcal{K}\mathcal{I}^{-1}, \tag{6}$$

where $\mathcal{I}$ is the information matrix based on the (non-estimated) log likelihood and $\mathcal{K}$ is the expected conditional variance of the non-validation contribution to the log likelihood (Pepe, 1992),

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

for parameters $\theta = \{F, G\}$. The first term in the Σ variance expression represents the variance component based on the maximum likelihood estimator and the second term represents a penalty from estimating the likelihood with empirical probabilities. The $\mathcal{I}$ and $\mathcal{K}$ matrices can be estimated consistently by

$$\widehat{\mathcal{I}} = \frac{1}{n}\frac{\partial^2 \log \widehat{L}}{\partial\theta^2}\bigg|_{\theta=\widehat{\theta}} \tag{7}$$

for maximum estimated likelihood estimates $\widehat{\theta} = \{\widehat{F}, \widehat{G}\}$ and

$$\widehat{\mathcal{K}} = \frac{1}{n_V}\sum_{i\in V}\widehat{Q}_i\widehat{Q}_i^T\bigg|_{\theta=\widehat{\theta}}, \tag{8}$$

where

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

and

$$D(X_i, \delta_i) = \frac{\partial P(X_i, \delta_i)}{\partial\theta}$$

$$\widehat{D}(X_j^*, \delta_j^*) = \sum_{k=1}^{K}\sum_{\delta=0}^{1}\frac{\partial P(X, \delta)}{\partial\theta}\widehat{P}(X_j^*, \delta_j^*|X, \delta).$$

In practice, derivatives in the variance expression can be calculated numerically. We found that the numerical derivatives were sometimes unable to be computed or led to negative variances with data that had large amounts of missingness or large numbers of parameters to estimate. In these cases, bootstrapped variance estimates can be calculated or analytical forms of the derivatives should be used. Furthermore, in order to calculate point-wise confidence intervals using the proposed parameter estimates and variance estimates, a log transformation or arcsine-square root transformation may be used to ensure confidence limits are bounded by 0 and 1 (Borgan and Liestøl, 1990).

4. Simulations

Our proposed survival function estimator is motivated by the fact that true endpoints are missing for some subjects while uncertain endpoints are available for all subjects and carry useful information for survival function estimation. In order to assess the performance of our proposed survival function estimator, we conducted a series of simulation studies. We simulated the true event time from a discrete uniform distribution, $T \sim \mathrm{Unif}[1,8]$, where survival time can only take integer values, and assumed right censoring at $C = 7$. The uncertain time to event was calculated as $T^* = T + \epsilon$, where $\epsilon \sim \mathrm{Unif}[0, \zeta]$ and ϵ is independent of T. The maximum integer value of the discrete uniform distribution for ϵ was calculated as $\zeta = \left\lfloor\sqrt{63 \cdot (1 - \rho^2)/\rho^2 + 1} - 1\right\rfloor$, where $\lfloor a\rfloor$ represents the largest

integer not greater than a and ρ represents the correlation between T and T^*. The expression for ζ was computed using the definition of correlation between T and T^*, independence of ϵ and T, and variance expressions for T and ϵ. Mathematical details of the derivation can be found in Web Appendix B. We considered correlations of $\rho \in \{0.01, 0.25, 0.50, 0.75, 1\}$. We set the right-censoring time for the uncertain endpoint also at $C^* = 7$. To create a representative validation subsample, we simulated data missing completely at random (MCAR) by randomly selecting a proportion $r \in \{0.25, 0.50, 0.75\}$ of the sample to be missing true endpoints. We used total sample sizes of $n \in \{200, 500\}$ and conducted 500 simulation repetitions for each set of parameter values.

For each simulation, we used the proposed method to calculate survival function estimates at each observed time. We also calculated complete-case Kaplan–Meier estimates using only true endpoints in the validation set, the naïve Kaplan–Meier estimates using only uncertain endpoints from all subjects, and the true Kaplan–Meier estimates using true endpoints from all subjects (which would be unavailable in real data). For the proposed estimator, the complete-case Kaplan–Meier estimator, and the naïve Kaplan–Meier estimator, we calculated estimated bias (parameter estimate−true parameter values), observed sample standard deviations (SD), estimated standard errors ($\widehat{\text{SE}}$), relative efficiency (RE) compared to the true Kaplan–Meier estimator (where lower RE implies greater efficiency and RE equal to 1 implies optimal efficiency), mean squared error (MSE) estimates, and 95% coverage (Cov) using log transformed confidence intervals at each of the observed time points. Each statistic was then averaged over all time points. We note that for all simulations presented in Tables 1–3, the observed sample standard deviation corresponds well with the standard error estimates from the asymptotic theory for the proposed estimator.

Table 1 shows the results from the simulation study with type 1 censoring and $n = 200$. The proposed estimator behaves similarly to the complete-case Kaplan–Meier estimator in terms of bias. Both have little bias, whereas the naïve Kaplan–Meier estimator is heavily biased. When the proportion of missingness is low or moderate ($r = 0.25$ or $r = 0.50$), the RE of our proposed estimator is similar to that of the complete-case Kaplan–Meier estimator when correlation is low and improves until it reaches optimal efficiency with correlation of 1, which can be interpreted as the situation where the uncertain outcome has no measurement error. The MSE of the proposed estimator is also similar to then becomes smaller than the MSE of the complete-case Kaplan–Meier estimator as correlation increases, and it is consistently smaller than the MSE of the naïve Kaplan–Meier estimator. This demonstrates that using the internal validation subsample can reduce the bias of survival estimates compared to using only uncertain endpoints and that using uncertain endpoints in the non-validation subsample can improve efficiency compared to using only true endpoints. When the amount of missing true outcomes is high ($r = 0.75$), though, our proposed estimator is actually slightly less efficient than the complete-case Kaplan–Meier estimator at low correlations between outcomes.

We saw similar results for simulations with $n = 500$, as shown in Web Table 1. We also increased the proportion of censored subjects (results not shown) by setting an earlier censoring time for both endpoints and arrived at the same conclusions. Although we assumed only non-negative measurement error of the uncertain endpoint for our simulations to demonstrate the potentially large bias in the naïve estimator and to better control the correlation between outcomes, we also conducted simulations allowing for negative or positive measurement error and the results (not shown) for our estimator and the complete-case estimator are similar.

In addition, we tested the performance of our proposed method at smaller sample sizes, $n \in \{10, 20, 30, \ldots\}$, to determine an approximate threshold for the number of events per parameter, or events per variable (EPV), needed for accurate estimation. We calculated the EPV as the smallest number of events in the validation set divided by $\widetilde{K} \in \{4, 7, 10\}$, the number of parameters to estimate, such that average bias was less than 0.01 in magnitude and average coverage was between 93% and 97%. Through these simulation studies (Web Table 2), we found an EPV of at most 4.

To compare the efficiency between our proposed estimator and the complete-case Kaplan–Meier estimator over various amounts of missingness, we computed the relative efficiencies (averaged over times) at 5% increments of the percentage of missingness of true endpoints for correlations of $\rho \in \{0.25, 0.50, 0.75\}$ (Figure 1). For these simulations, we used a larger sample size of $n = 500$ to ensure that the EPV was adequate even at the largest amounts of missingness. For correlations of 0.50 and 0.75, our proposed estimator is more efficient (lower RE) than the complete-case Kaplan–Meier estimator when the proportion of missing data is low, then the efficiency curves cross and our proposed estimator becomes less efficient. The point of crossing is at a higher percentage of missingness with higher values of the correlation between outcomes. Even with low correlation ($\rho = 0.25$) between outcomes, though, our estimator has similar or lower efficiency than the complete-case Kaplan–Meier estimator when the amount of missingness is 50% or less. This is consistent with Pepe's recommendation for non-survival data with one parameter of interest (Pepe, 1992).

We explored the behavior of our proposed estimator under random censorship by simulating true event times $T \sim$Unif$[1, 8]$, uncertain event times $T^* = T + \epsilon$ where $\epsilon \sim$Unif$[0, 2]$, true censoring times $C \sim$Unif$[5, 7]$, and uncertain censoring times $C^* = C + \gamma$ where $\gamma \sim$Unif$[0, 2]$. These simulations resulted in a small amount of censoring (approximately 25%). We also increased the amount of censoring by simulating true censoring times $C \sim$Unif$[3, 7]$, which resulted in a moderate amount of censoring (approximately 38%). The results of these random censoring simulations are shown in Table 2. Similar to the results from type 1 censoring, our proposed estimator has little bias compared to the naïve Kaplan–Meier estimator and is more efficient than the complete-case Kaplan–Meier estimator for both small and moderate amounts of censoring. We saw similar results with $n = 500$ as seen in Web Table 3.

To test the robustness of the MCAR assumption of the proposed method, we relaxed this assumption and simulated data missing at random (MAR). We defined a missingness indicator R, where $R = 1$ denotes a missing true endpoint and $R = 0$ denotes a non-missing true endpoint, based on the uncertain

Table 1
Simulation results for type 1 censoring and $n = 200$

r	ρ	Method	Bias $\times 10^{-3}$	SD	$\widehat{\text{SE}}$	MSE $\times 10^{-3}$	RE	Cov
		Proposed	−0.26	0.035	0.035	1.27	1.36	0.96
	0.01	Comp K–M	−0.55	0.035	0.035	1.27	1.36	0.95
		Naïve K–M	498.41	0.003	0.002	310.32	0.01	0.00
		Proposed	0.79	0.035	0.035	1.28	1.36	0.96
	0.25	Comp K–M	−0.55	0.035	0.035	1.27	1.36	0.95
		Naïve K–M	449.43	0.014	0.014	247.32	0.28	0.00
		Proposed	0.57	0.035	0.034	1.26	1.34	0.96
25	0.50	Comp K–M	−0.55	0.035	0.035	1.27	1.36	0.95
		Naïve K–M	384.50	0.020	0.020	175.53	0.56	0.00
		Proposed	0.32	0.034	0.033	1.18	1.26	0.96
	0.75	Comp K–M	−0.55	0.035	0.035	1.27	1.36	0.95
		Naïve K–M	285.90	0.025	0.025	91.52	0.83	0.00
		Proposed	0.02	0.030	0.030	0.94	1.00	0.96
	1.00	Comp K–M	−0.55	0.035	0.035	1.27	1.36	0.95
		Naïve K–M	0.03	0.030	0.030	0.94	1.00	0.95
		Proposed	−0.15	0.043	0.043	1.88	1.99	0.95
	0.01	Comp K–M	−1.01	0.043	0.042	1.87	1.98	0.95
		Naïve K–M	498.41	0.003	0.002	310.32	0.01	0.00
		Proposed	4.62	0.044	0.042	2.02	2.14	0.95
	0.25	Comp K–M	−1.01	0.043	0.042	1.87	1.98	0.95
		Naïve K–M	449.43	0.014	0.014	247.32	0.28	0.00
		Proposed	3.10	0.044	0.042	1.94	2.05	0.95
50	0.50	Comp K–M	−1.01	0.043	0.042	1.87	1.98	0.95
		Naïve K–M	384.50	0.020	0.020	175.53	0.56	0.00
		Proposed	2.32	0.042	0.040	1.76	1.88	0.95
	0.75	Comp K–M	−1.01	0.043	0.042	1.87	1.98	0.95
		Naïve K–M	285.90	0.025	0.025	91.52	0.83	0.00
		Proposed	0.02	0.030	0.030	0.94	1.00	0.96
	1.00	Comp K–M	−1.01	0.043	0.042	1.87	1.98	0.95
		Naïve K–M	0.03	0.030	0.030	0.94	1.00	0.95
		Proposed	3.35	0.061	0.060	3.86	4.11	0.96
	0.01	Comp K–M	1.86	0.061	0.060	3.83	4.07	0.96
		Naïve K–M	498.41	0.003	0.002	310.32	0.01	0.00
		Proposed	24.12	0.065	0.065	4.39	4.65	0.98
	0.25	Comp K–M	1.86	0.061	0.060	3.83	4.07	0.96
		Naïve K–M	449.43	0.014	0.014	247.32	0.28	0.00
		Proposed	21.49	0.067	0.063	4.58	4.83	0.96
75	0.50	Comp K–M	1.86	0.061	0.060	3.83	4.07	0.96
		Naïve K–M	384.50	0.020	0.020	175.53	0.56	0.00
		Proposed	9.84	0.061	0.058	3.83	4.13	0.95
	0.75	Comp K–M	1.86	0.061	0.060	3.83	4.07	0.96
		Naïve K–M	285.90	0.025	0.025	91.52	0.83	0.00
		Proposed	−0.22	0.031	0.030	0.97	1.03	0.96
	1.00	Comp K–M	1.86	0.061	0.060	3.83	4.07	0.96
		Naïve K–M	0.03	0.030	0.030	0.94	1.00	0.95

r is the percent missing and ρ is the correlation between true and uncertain outcomes. Proposed refers to the proposed estimator, Comp K–M refers to the complete-case Kaplan–Meier estimator, and naïve K–M refers to the naïve Kaplan–Meier estimator. SD is standard deviation of estimates across simulations, $\widehat{\text{SE}}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

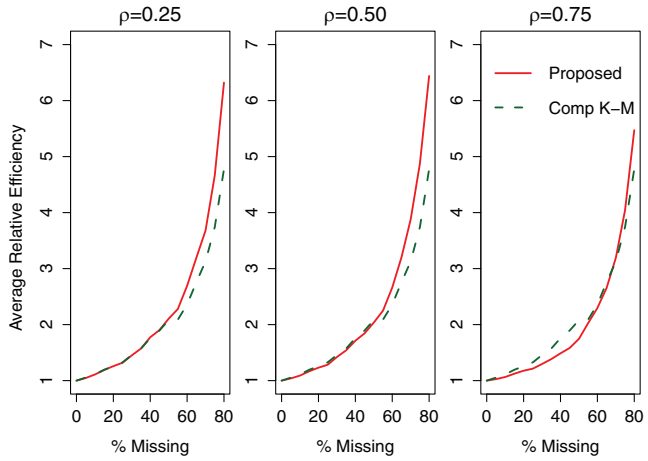

Figure 1. Relative efficiencies by correlation between true and uncertain endpoints (ρ) and amount of missingness of true endpoints. Proposed refers to the proposed estimator and Comp K–M refers to the complete-case Kaplan–Meier estimator.

indicator δ^* such that

$$R|(\delta^* = 0) = \begin{cases} 1 \text{ with probability } p_R \\ 0 \text{ with probability } 1 - p_R \end{cases}$$

$$R|(\delta^* = 1) = \begin{cases} 1 \text{ with probability } 1 - p_R \\ 0 \text{ with probability } p_R \end{cases}$$

for probability $p_R = 0.60$. This implies that the probability of missingness of the true endpoint depends on the observed event indicator of the uncertain endpoint. In the AD example, this would mean that subjects who are clinically determined to be non-AD during the study are more likely to miss the CSF biomarker endpoint. In the results from the MAR simulations in Table 3, we see that both the proposed estimator and the complete-case Kaplan–Meier estimator are sometimes slightly biased. However, the proposed estimator is less biased than the complete-case estimator, particularly when the correlation between outcomes is very high, and therefore the coverage of the proposed estimator is better than the coverage of the complete-case estimator when the correlation between outcomes is greater than 0.01. We saw similar results with $n = 500$ as seen in Web Table 4.

Table 2
Simulation results for random censoring and $n = 200$

			Bias			MSE		
r	C	Method	$\times 10^{-3}$	SD	$\widehat{\text{SE}}$	$\times 10^{-3}$	RE	Cov
25	S	Proposed	0.24	0.035	0.034	1.25	1.17	0.96
		Comp K–M	0.39	0.038	0.037	1.47	1.39	0.95
		Naïve K–M	119.63	0.030	0.029	15.47	0.83	0.03
	M	Proposed	0.57	0.040	0.038	1.65	1.20	0.96
		Comp K–M	0.37	0.043	0.041	1.94	1.44	0.94
		Naïve K–M	119.98	0.032	0.032	15.75	0.78	0.05
50	S	Proposed	−2.18	0.040	0.041	1.64	1.54	0.95
		Comp K–M	−0.12	0.047	0.046	2.23	2.10	0.95
		Naïve K–M	119.63	0.03	0.029	15.47	0.83	0.03
	M	Proposed	−1.99	0.049	0.044	2.68	1.84	0.96
		Comp K–M	−0.52	0.053	0.050	2.95	2.18	0.93
		Naïve K–M	119.98	0.032	0.032	15.75	0.78	0.05

r is the percent missing and C is the amount of censoring, where S means small (25%) and M means moderate (38%). Proposed refers to the proposed estimator, Comp K–M refers to the complete-case Kaplan–Meier estimator, and Naïve K–M refers to the naïve Kaplan–Meier estimator. SD is standard deviation of estimates across simulations, $\widehat{\text{SE}}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

5. Application to the Alzheimer's Disease Neuroimaging Initiative Study

We illustrated our method by considering data (retrieved on July 26, 2013) from the ongoing ADNI study (Weiner et al., 2012). See Web Appendix C for more detailed information about the ADNI study. Participants in this study were seen every 6 months until the end of 2 years, then annually thereafter, at which time clinical diagnoses of non-AD (cognitively normal or mild cognitive impairment) or AD were given. These follow-up times were predetermined by study design, and AD is a chronic disease with slow progression (Jack et al., 2010). Therefore, the outcome of interest in the current study was time to detection of AD, measured in years, and thus discrete survival estimates would be appropriate. The current study includes data from participants in the ADNI-1 and ADNI-GO segments of the ADNI study. For those who agreed to a lumbar puncture, CSF assays were performed and $A\beta$ protein concentrations were measured. Participants with an $A\beta$ biomarker value greater than 192 pg/ml were classified as non-AD at baseline and those with an $A\beta$ value less than or equal to 192 pg/ml were classified as AD at baseline (Shaw et al., 2009). There were 186 patients who were non-AD at the time of enrollment according to both the clinical diagnosis and CSF diagnosis. For each patient, the time to clinical AD or last follow-up was recorded to obtain an uncertain, mismeasured outcome on all patients. A subset of 110 patients continued to have CSF assays performed annually. For these 110 patients in the validation set, patients were classified as non-AD or AD at each time point using the same cutoff of 192 pg/ml and the true time to AD or last follow-up was also recorded. Thus, patients with any CSF assays during follow-up were considered to be in the validation set and those with no follow-up CSF assays were in the non-validation set, or $n_V = 110$ and $n = 186$ using the notation of Section 2.

First, we assessed the missingness in the data. We used a log-rank test to compare the survival functions for time to clinical AD between the non-validation set and the validation set. The χ^2 test statistic was 0.2 with 1 degree of freedom, yielding a p-value of 0.662. We also used Fisher's exact test to test for an association between the clinical event indicator and missingness. The p-value was 1. Further, because we used all available longitudinal CSF assays, those who were missing CSF outcomes were missing immediately after baseline. Since all subjects begin as non-AD, the missingness could not be dependent on baseline CSF or clinical diagnoses. Therefore,

Table 3
Simulation results for data missing at random and $n = 200$

Censoring	ρ/C	Method	Bias $\times 10^{-3}$	SD	$\widehat{\text{SE}}$	MSE $\times 10^{-3}$	RE	Cov
Type 1	0.01	Proposed	2.53	0.048	0.048	2.31	2.47	0.96
		Comp K–M	0.91	0.048	0.048	2.31	2.47	0.95
		Naïve K–M	498.41	0.003	0.002	310.32	0.01	0.00
	0.25	Proposed	17.06	0.048	0.049	2.38	2.56	0.97
		Comp K–M	−11.75	0.046	0.046	2.17	2.31	0.94
		Naïve K–M	449.43	0.014	0.014	247.32	0.28	0.00
	0.50	Proposed	21.47	0.047	0.046	2.22	2.38	0.98
		Comp K–M	−26.55	0.045	0.045	2.06	2.19	0.90
		Naïve K–M	384.50	0.020	0.020	175.53	0.56	0.00
	0.75	Proposed	16.27	0.042	0.041	1.78	1.94	0.96
		Comp K–M	−44.30	0.043	0.042	1.89	2.00	0.81
		Naïve K–M	285.90	0.025	0.025	91.52	0.83	0.00
	1.00	Proposed	0.02	0.030	0.030	0.94	1.00	0.96
		Comp K–M	−22.56	0.039	0.039	1.57	1.65	0.88
		Naïve K–M	0.03	0.030	0.030	0.94	1.00	0.95
Random	S	Proposed	0.32	0.039	0.044	1.62	1.49	0.96
		Comp K–M	−33.92	0.043	0.042	1.87	1.78	0.86
		Naïve K–M	119.63	0.03	0.029	15.47	0.83	0.03
	M	Proposed	1.66	0.047	0.044	2.49	1.69	0.96
		Comp K–M	−41.74	0.048	0.047	2.31	1.81	0.83
		Naïve K–M	119.98	0.032	0.032	15.75	0.78	0.05

Censoring is the type of the censoring mechanism and ρ/C either represents the correlation ρ between true and uncertain outcomes or represents the amount of censoring, where S means small (25%) and M means moderate (38%). Proposed refers to the proposed estimator, Comp K–M refers to the complete-case Kaplan–Meier estimator, and naïve K–M refers to the naïve Kaplan–Meier estimator. SD is standard deviation of estimates across simulations, $\widehat{\text{SE}}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

we did not find strong evidence against the MCAR assumption.

Figure 2 shows the estimated survival functions using our proposed estimator, which maximized the estimated likelihood, the complete-case Kaplan–Meier estimator which only uses 110 CSF diagnoses, and the naïve Kaplan–Meier estimator which only uses the 186 clinical diagnoses. The three survival functions are very similar until 36 months, at which time the naïve Kaplan–Meier estimate begins to diverge from the other two survival curves. With higher survival probabilities, the naïve estimate overestimates the probability of being AD-free after 36 months compared to the proposed estimator and complete-case Kaplan–Meier estimator. Since the naïve estimate is based on only clinical diagnoses, this would indicate that abnormality of $A\beta$ occurred earlier than cognitive impairment. This finding is consistent with the recent theoretical model of AD pathology developed by Jack et al.(2010).

Table 4 shows the standard error estimates at each time point. The standard errors of the proposed estimate are similar to or smaller than those of the complete-case Kaplan–Meier estimate at all time points. Although we cannot calculate the correlation between true and uncertain event times in the data example since both events are not observed for any subjects, the estimated correlation between the true and uncertain observed outcomes is 0.363 (Hotelling, 1936). Thus, the standard errors we observed further support the conclusion that the proposed estimator helps to improve efficiency relative to the complete-case estimator when there is moderate correlation between outcomes.

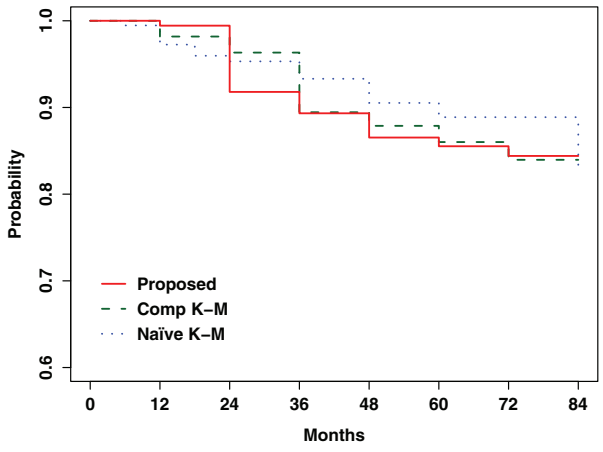

Figure 2. Data example survival function estimates for time to AD. Proposed refers to the proposed estimator, Comp K–M refers to the complete-case Kaplan–Meier estimator, and Naïve K–M refers to the naïve Kaplan–Meier estimator.

Table 4
Data example standard error estimates

Month	Proposed estimator	Complete-case Kaplan–Meier	Naïve Kaplan–Meier
6	0.000	0.000	0.005
12	0.008	0.013	0.012
18	0.008	0.013	0.016
24	0.022	0.019	0.017
36	0.036	0.036	0.023
48	0.038	0.040	0.031
60	0.040	0.045	0.036
72	0.046	0.051	0.036
84	0.046	0.051	0.074

6. Discussion

We proposed a nonparametric maximum likelihood estimator for the discrete survival function in the presence of uncertain endpoints by using an internal validation subsample. We allowed for random censoring for survival outcomes by incorporating a censoring distribution in the likelihood, showed that the survival function estimator is a step function that drops only at observed event times, and proved that the proposed estimator is consistent and asymptotically normal. We evaluated the finite sample performance of the proposed estimator through extensive simulations. We found that the proposed estimator has little bias and can improve efficiency relative to the complete-case Kaplan–Meier estimator. It can also reduce bias compared to the naïve Kaplan–Meier estimator. The proposed estimator also works better than the complete-case and naïve estimators under departure from the MCAR assumption.

The efficiency gains of the proposed estimator have useful implications in clinical trials. A true outcome may be costly to obtain on all subjects, but using the proposed method can incorporate a less costly uncertain outcome assessed on all subjects and the true outcomes on a smaller subsample. Compared to obtaining true outcomes on all subjects which can be very costly or using a complete-case estimator on the smaller subsample, our estimator can reduce costs of a trial without sacrificing power.

The proposed approach does not require that subjects can only be validated at the end of study. Instead, our approach allows that all subjects can have the opportunity to be validated at any predetermined timepoint. Through simulations, we found that the efficiency gains of our proposed estimator depends on both the correlation between outcomes and the size of the validation sample. However, in general, the proposed estimator appears to work well when the size of the validation sample is 50% or more of the total sample size. Interesting study design issues arise with regard to the size of the validation sample that is needed to adequately accommodate the uncertainty of the mismeasured endpoints. For example, in epidemiological studies, if the total budget for assessing subjects is fixed, it is valuable to determine the optimum size of the study cohort and optimum size of the validation subsample. These design issues are currently under investigation.

The proposed method can be used with data that have both type 1 right censoring and random right censoring. The proposed method also assumes that study subjects are seen at predetermined time points and relies on a discrete time framework. In studies where subjects are evaluated at any time, that is, time is considered continuous, the proposed estimator may not improve efficiency compared to the complete-case Kaplan–Meier estimator. Furthermore, if study participants are not seen as frequently as the unit of time of interest or if a participant misses a visit and then experiences the event upon return, there may be interval censoring. For these situations, an extension of our proposed method is needed, but is not trivial.

Currently, our proposed method only estimates a single survival function. A natural extension of the method would be a semiparametric version that is able to incorporate covariates and conduct between-group comparisons. Using our proposed method for a proportional hazards model to estimate a hazard ratio for a categorical or continuous covariate of interest is currently under investigation. The case of a categorical covariate is similar to the nonparametric version, but with the addition of a random variable in the estimated likelihood. For the continuous covariate, however, a more complex approach such as inclusion of smooth kernel-type functions in the empirical probability estimates must be used.

As early detection of AD and other chronic diseases becomes increasingly important, but event outcomes may be hard to obtain for everyone, we recommend collecting an internal validation sample when the measures of the event outcome are uncertain so that statistical analysis can be improved with greater accuracy and power.

7. Supplementary Materials

Web appendices, including detailed mathematical derivations, additional information about the ADNI study, and additional results from simulation studies referenced in Sections 2, 4, and 5, as well as R code for implementing the proposed method, are available with this paper at the *Biometrics* website on the Wiley Online Library.

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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. 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. The authors thank an anonymous associate editor and two anonymous referees for their constructive comments.

References

Balasubramanian, R. and Lagakos, S. W. (2001). Estimation of the timing of perinatal transmission of HIV. *Biometrics* **57**, 1048–1058.

Borgan, O. R. and Liestøl, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scandinavian Journal of Statistics* **17**, 35–41.

Fleming, T. R., Prentice, R. L., Pepe, M. S., and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* **13**, 955–968.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **3/4**, 321–377.

Jack, C. R. Jr., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurology* **9**, 119–128.

Magaret, A. S. (2008). Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Statistics in Medicine* **27**, 5456–5470.

Meier, A. S., Richardson, B. A., and Hughes, J. P. (2003). Discrete proportional hazards models for mismeasured outcomes. *Biometrics* **59**, 947–954.

Nelson, P. T., Alafuzoff, I., Bigio, E. H., Bouras, C., Braak, H., Cairns, N. J., et al. (2012). Correlation of Alzheimer disease neuropathologic changes with cognitive status: A review of the literature. *Journal of Neuropathology and Experimental Neurology* **71**, 362–381.

Pepe, M. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–365.

Richardson, B. A. and Hughes, J. P. (2000). Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics* **1**, 341–354.

Shaw, L., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., et al. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology* **65**, 403–413.

Snapinn, S. M. (1998). Survival analysis with uncertain endpoints. *Biometrics* **54**, 209–218.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2012). The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimer's & Dementia* **8**, S1–68.

Received October 2013. *Revised February* 2015. *Accepted March* 2015.