

Nonparametric Survival Estimation when Death is Reported with Delay

ALAN E. HUBBARD¹

MARK J. VAN DER LAAN¹

WAYNE ENANORIA²

JOHN M. COLFORD, JR. ²

University of California, School of Public Health, Division of Biostatistics¹, Division of Public Health Biology and Epidemiology², Berkeley, CA

Abstract. In disease registries there can be a delay between death of a subject and the reporting of this death to the data analyst. If researchers use the Kaplan-Meier estimator and implicitly assumed that subjects who have yet to have death reported are still alive, i.e. are censored at the time of analysis, the Kaplan-Meier estimator is typically inconsistent. Assuming censoring is independent of failure, we provide a simple estimator that is consistent and asymptotically efficient. We also provide estimates of the asymptotic variance of our estimator and simulations that demonstrate the favorable performance of these estimators. Finally, we demonstrate our methods by analyzing AIDS survival data. This analysis underscores the pitfalls of not accounting for delay when estimating the survival distribution and suggests a significant reduction in bias by using our estimator.

Keywords: Right-censored data, reporting delays, influence curve, Kaplan-Meier estimator

1 Introduction

1.1 Background

To study survival patterns in a population, health agencies will sometimes create large registries of individuals who have been diagnosed with a disease of interest, such as AIDS. To determine the survival distribution of, for instance, time from diagnosis of a disease to death, researchers cross-reference their registry with databases that contain the date of death for a larger population. Frequently, the recording or reporting of the death of an individual in the registry is done with some delay. An example of this type of data is the AIDS registry maintained by The Office of AIDS of the California Department of Health Services in cooperation with the Centers for Disease Control Prevention (see Colford, et al., 1997). This registry contains, in addition to several demographic variables, the date of AIDS diagnosis for all subjects and date of death and date of death report for those subjects who have had their death reported. This is an on-going registry and periodically the data is examined for estimation of the survival distribution by particular cohorts (defined by date of diagnosis). Note, the analyst does not know whether or not the subjects who have yet to have their death reported are still alive because their death report can be delayed.

This type of data is common and previous work includes methods to estimate the delay distribution. This is done to account for the under-reporting due to delay when attempting to estimate the number of deaths from the disease at a particular time (Bacchetti, 1996). In addition, Tu, et al. (1993) presented a estimator that relies on modeling of the delay distribution and imputation of the number of unreported deaths at a discrete time t . Hu and Tsiatis (1996) discuss estimation of the survival distribution from data more typical of clinical settings. Specifically, the data structure Hu and Tsiatis base their estimators on

contains several monitoring times when the vital status of the patient is reported, possibly with delay. Finally, Van der Laan and Hubbard (1998) extend this data structure to include time-dependent covariates and present a locally efficient one-step estimator that utilizes the covariates to account for dependent censoring and increase efficiency.

Let T be the failure time of interest (in the example above, T is the time from AIDS diagnosis to death). Define C to be the length of time from the index date to end of follow-up (in the above example, C is the time from AIDS diagnosis to time of analysis or last data collection). Finally, define V to be the time from the index date to the date the event of interest is reported (for the AIDS example, this is the time from AIDS diagnosis to report of death). To estimate the survival distribution, the analyst might typically define the failure time as T , the censoring time as C and use a Kaplan-Meier estimate on the data $(\tilde{T} = \min(T, C), \Delta = I(T \leq C))$. If some of the subjects have death reported with delay, i.e. $V \neq T$, then by doing the naive Kaplan-Meier estimator as above, one will be wrongly assuming for some subjects that $T > C$, when in fact $T \leq C$, but $V > C$. The effect will be to over-estimate the survival distribution, and this bias increases as more subjects are censored and the delay between death and death report increase. If analysis time is fixed and one applies the Kaplan-Meier to progressively younger cohorts, then, due to increased censoring, more recent cohorts will appear to have longer survival even if no real increase has occurred. In this paper, we present an estimator that efficiently estimates the survival distribution when death is reported with delay.

1.2 The data structure and model

Let the full data be $X = (T, V)$ and the observed data $Y = (\tilde{T} = \min(V, C), \Delta = I(V \leq C), \Delta T)$, where ΔT refers to the fact that the analyst only observes T if $\Delta = 1$. The analyst

observes n independent and identically distributed observations Y_1, \dots, Y_n of Y .

Since Y is a function of X and C , its distribution is indexed by the distribution of X and the conditional distribution of C , given X . The distribution F_X of X will be completely unspecified and we assume that the conditional distribution $G(\cdot | X)$ of C , given X , satisfies:

$$G(c | X) = G(c). \tag{1}$$

We note that our estimate of $F(t)$ requires that

$$\frac{I(T \leq t)}{\bar{G}(V)} > 0 \text{ } F_X \text{ almost everywhere,} \tag{2}$$

where $\bar{G}(c) = P(C \geq c)$. We discuss this important assumption and consequences of its violation in the last section.

The estimators we discuss can be extended to incorporate baseline and time-dependent covariates. However, we restrict our attention to the simpler setting as it is more typical of surveillance data. Van der Laan and Hubbard (1998) discuss locally efficient one-step estimators for the survival distribution when the data includes a delay/monitoring process and possibly both baseline and time-dependent covariates. This paper examines in detail a special case of the more general data structure presented in their paper. Specifically, we discuss when 1) $V_1(t)$, the general delay-in-reporting process, jumps only at V , the time of death report ($V_1(t)$ is the first time the status of the patient is known at time t), and 2) no informative covariates exist. We demonstrate that the simple estimator that van der Laan and Hubbard (1998) present is actually efficient for our data structure, so no one-step correction is needed.

1.3 Organization of Paper

First, we discuss our “inverse probability of censoring weighted” (IPCW) estimator (this terminology is introduced in Robins and Rotnitzky, 1992) of $F(t)$. This estimator is a mod-

ification of the Kaplan-Meier estimator since it reduces to the Kaplan-Meier estimator with probability one if $V = T$. We also provide a method for estimating the asymptotic variance of this estimator. By presenting the data as a right-censored estimation problem of the type originally studied by Robins and Rotnitzky (1992) and Robins (1993), and later adapted for nonparametric estimation of the survival distribution by Hubbard et al. (1998), we can use previous general results to prove that this estimator is semiparametrically efficient. We present a simulation study comparing the naive Kaplan-Meier to our IPCW estimator. The simulation is constructed to mimic the type of AIDS surveillance data discussed in the introduction. We use our proposed estimators to estimate the survival distribution (and corresponding quantiles) of the AIDS registry data and conclude the paper with a discussion regarding the possibility of very long delay times that violate assumption (2).

2 IPCW Estimator

We define the same estimator presented in van der Laan and Hubbard (1998) that weights the observed $I(T_i \leq t)$ by the correct probability of censoring. We exploit the following key identity to construct the IPCW estimator, given (2),

$$E \left\{ \frac{I(T \leq t)\Delta}{\bar{G}(\tilde{T} | X)} \right\} = F(t), \quad (3)$$

where $\bar{G}(c | X) = \bar{G}(c)$. This identity follows directly from (1),

$$E(\Delta | X) = P(C \geq V | X) = \bar{G}(V),$$

which shows that the conditional expectation given X of the left-hand side of (3) equals $I(T \leq t)$. This leads to the following estimator:

$$F_n^0(t) = \frac{1}{n} \sum_{i=1}^n \frac{I(T_i \leq t)\Delta_i}{\bar{G}_n(\tilde{T}_i)}, \quad (4)$$

where $\bar{G}_n(\cdot)$ is the Kaplan-Meier estimator of \bar{G} based on the n observations of $(\tilde{T} = C \wedge V, 1 - \Delta)$, where V now plays the role of the censoring variable for C . In the case of no delay, that is $V = T$, then this estimator equals the Kaplan-Meier estimator of $F(t)$, which is of course consistent if there is no delay between death and its recording in the database. Thus, this IPCW estimator is a generalization of the Kaplan-Meier estimator.

In the theorem below, we prove that this estimator is asymptotically efficient. For efficiency theory we refer to Bickel, et al. (1993). For our purposes, it is sufficient to note that an estimator $F_n(t)$ of $F(t)$ is asymptotically linear with influence curve $IC(Y | F_X, F(t), G)$ if

$$F_n(t) = F(t) + \frac{1}{n} \sum_{i=1}^n IC(Y_i | F_X, F(t), G) + op(1/\sqrt{n}),$$

and it is asymptotically efficient if the influence curve equals the so called efficient influence curve.

Theorem: *Given (2) then (4) is asymptotically efficient with influence curve,*

$$IC^*(Y | F_X, F(t), G) = \frac{\Delta I(T \leq t)}{\bar{G}(V)} - F(t) + \int F(t | \tilde{T} > u) \frac{dM(u)}{\bar{G}(u)}. \quad (5)$$

Proof of Theorem

To show that (5) is the efficient influence curve, one can use the efficient influence curve presented in van der Laan and Hubbard (1998) and apply it to our special case. However, we present more detail to give the reader insight into the proof. Specifically, to find the influence curve for the IPCW estimator with no covariates, we represent the delay data as right-censored data and then we can directly apply results from Robins and Rotnitzky (1992). Let t be given. We define a process $V_1(u)$, which is equal to T only when $u \geq V$ and is 0 otherwise. $V_1(u)$ gives the state of knowledge about the failure time of the individual at time

u . Also, define $\bar{V}_1(u) = \{V_1(c) : c < u\}$. Then, the observed data, Y , can be written as,

$$Y = (\tilde{T} = \min(V, C), \Delta = I(V \leq C), \bar{V}_1(\tilde{T})). \quad (6)$$

By Robins and Rotnitzky (1992), the efficient influence curve, IC^* , for estimation of $F(t)$ with right-censored data (6) is:

$$IC^*(Y | F_X, F(t), G) = IC_0(Y | F(t), G) - \Pi(IC_0(Y | F(t), G) | T_{CAR}), \quad (7)$$

where $IC_0 = \Delta I(T \leq t) / \bar{G}(T | X) - F(t)$ and $\Pi(IC_0(Y | F(t), G) | T_{CAR})$ is the Hilbert space projection of IC_0 onto the tangent space, T_{CAR} , of scores of the nuisance parameter G only assuming (1). Here we use that T_{CAR} equals the orthogonal complement of the tangent space for F_X (Gill et al., 1998). The space T_{CAR} is given by (Robins and Rotnitzky, 1992):

$$T_{CAR} = \left\{ \int h(u, \bar{V}_1(u)) dM(u) : h \right\},$$

where

$$dM(u) \equiv I(C \in du, \Delta = 0) - \Lambda_C(du | X) I(\tilde{T} > u). \quad (8)$$

In our case, because $V_1(u) = 0$ for $u < V$, and so $V_1(C) = 0$ if $C < V$, we can represent T_{CAR} as,

$$T_{CAR} = \left\{ \int h(u) dM(u) : h \right\}.$$

The projection $IC_{nu}^* \equiv \Pi(IC_0(Y | F(t), G) | T_{CAR})$ is given by

$$IC_{nu}^*(Y | F_X, G) = - \int F(t | \bar{V}_1(u), \tilde{T} > u) \frac{dM(u)}{\bar{G}(u | X)}, \quad (9)$$

where $F(t | \bar{V}_1(u), \tilde{T} > u)$ is the conditional probability that $T \leq t$, given $\bar{V}_1(u)$ and $\tilde{T} > u$.

Thus, the efficient influence curve can be written as,

$$IC^*(Y | F_X, F(t), G) = IC_0(Y | F(t), G) - IC_{nu}^*(Y | F_X, G). \quad (10)$$

If we can show that the efficient influence curve of our estimator is equivalent to (10), then our estimator is efficient (Bickel, et al., 1993). For our problem, note that $\bar{V}_1(u) = 0$ if $u < V$ so $\bar{V}_1(u)$ adds no information about T if $u < V$. Thus, one can replace $F(t | \bar{V}_1(u), \tilde{T} > u) = F(t | \tilde{T} > u)$ in (9) and IC^* can be expressed explicitly as (5).

It is shown in Robins and Rotnitzky (1992) that if 1) G is estimated efficiently assuming the given independent censoring model for G , 2) the IPCW estimator is asymptotically linear and 3) assumption (2) holds, then the IPCW estimator has influence curve,

$$IC(Y | F_X, F(t), G) = IC_0(Y | F(t), G) - \Pi(IC_0(Y | F(t), G) | T_2), \quad (11)$$

where T_2 is the tangent space of the scores for estimation of G assuming independent censoring. Since G is estimated efficiently with the Kaplan-Meier estimator, then the IPCW estimator has influence curve (11). Finally, because Robins (1996) demonstrated that $T_2 = \{ \int h(u) dM(u) : h \}$ and so in this case $T_2 = T_{CAR}$, by (7) the influence curve of our IPCW estimator is equal to the efficient influence curve. Thus, the IPCW estimator with no covariates and G estimated with the Kaplan-Meier estimator is asymptotically efficient with influence curve (5). \square

2.1 Estimating the Variance

In order to estimate the variance of the IPCW estimator, we need an estimate of this influence curve. We estimate this influence curve with $IC(Y | F_{X,n}, F_n^0(t), G_n)$ obtained by substitution of an estimator $F_n(t | \tilde{T} > u)$ of $F(t | \tilde{T} > u)$, the estimator $F_n^0(t)$ (4) and G_n . Since G_n is the Kaplan-Meier estimator, the integral on the right side of (5) jumps only at the censoring times, i.e., it is a sum. Thus, one only needs to estimate $F_n(t | \tilde{T} > u)$ at u 's corresponding to the observed censoring times. To do this, one simply performs the IPCW estimator of $F(t)$ (4) on the subsample $\tilde{T} > u$ for u 's that consist of all of the observed censoring times. Of

course, typically one will want to estimate $F(t)$ for a large set of t 's, and so both the estimator and its variance estimate will have to be repeated for each t . The estimated influence curve can be written explicitly as:

$$IC(Y_i | F_{X,n}, F_n^0(t), G_n) = \frac{\Delta_i I(T_i \leq t)}{\bar{G}_n(T_i)} - F_n^0(t) + \left[\frac{(1 - \Delta_i) F_n(t | \tilde{T} > C_i)}{\bar{G}_n(C_i)} - \sum_{u_j < \tilde{T}_i} F_n(t | \tilde{T} > u_j) \frac{\lambda_n(u_j)}{\bar{G}_n(u_j)} \right], \quad (12)$$

where u_j is the j th ordered observed censoring time and $\lambda_n(u_j)$ is the corresponding Kaplan-Meier estimate of the hazard. Finally, an estimate of the asymptotic variance of $F_n^0(t)$ is given by,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n IC^2(Y_i | F_{X,n}, F_n^0(t), G_n).$$

This variance estimate can be used to construct a $1 - \alpha$ confidence interval,

$$F_n^0(t) \pm z_{1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}. \quad (13)$$

3 Simulation Results

In this section, we examine the finite-sample-performance of the IPCW estimator relative to the Kaplan-Meier estimator. We chose to do present just the results of a single simulation for each data-generating distribution because the advantage of this estimator at the simulation sample size ($n=1000$) is from a reduction in bias, and this can be seen clearly in a single simulation (see van der Laan and Hubbard, 1998 for additional repeated simulations that also include informative covariates). The data-generating distributions were chosen to mimic the AIDS survival data we examine in the next section. The time of analysis was fixed at t_a , time of AIDS diagnosis follows a uniform $(0, t_d)$ distribution, time of death, T ,

comes from an exponential (λ_T) distribution, and $V = T + U$, where $U = \min(C_0, Q)$ and $Q \sim \text{exponential}(\lambda_Q)$. In this case, the censoring time, C is the time from AIDS diagnosis to time of analysis, and subjects are not included if time of diagnosis is after time of analysis. We did two simulations, both of sample size of 1000. In the first simulation (A), $t_a = 5$, $t_d = 3$, $\lambda_T = 1$, $\lambda_Q = 2.0$ and $C_0 = 1.0$. This results in data with delay, but few censored observations. In this case, the Kaplan-Meier estimator should be a reasonably consistent estimate of the survival distribution. In the second simulation (B), $t_a = 2$, $t_d = 3$, $\lambda_T = 1$, $\lambda_Q = 0.5$ and $C_0 = 1.0$. Because the time of analysis is early relative to the distribution of failure times, many of the observations are censored. In this case, the Kaplan-Meier estimator should be significantly (and positively) biased, but our IPCW estimator should remain consistent. The results of these simulations are shown in figure 1. In addition, we show the pointwise confidence intervals for simulation B using both our estimator using (13) and that using Greenwood's formula for the Kaplan-Meier estimator (see figure 2).

The results confirm expectations. When there are relatively few censored observations (time of analysis is large relative to the failure times) then the delay has little effect. However, as relatively more recent cohorts are analyzed, then the sample will contain more censored observations and the Kaplan-Meier estimator becomes progressively more biased. In addition, note that the confidence intervals are of similar widths when the IPCW and Kaplan-Meier estimates are similar. If no delay occurs, then the influence curve of our estimator reduces to that of the Kaplan-Meier estimator and so not only will the IPCW and the Kaplan-Meier estimators be equivalent, but the estimates of their variances will also be equivalent. In the next section, we analyze a data set that inspired the above simulations.

4 Data Analysis

In this section we apply our IPCW-estimator (4) to a data set that contains for each subject the time death of the subject (from diagnosis of AIDS) and the time of death report for those subjects who have a reported death before analysis time, and simply follow up time for the remaining subjects. The data comes from a registry managed by the Office of AIDS of the California Department of Health Services, in cooperation with the Centers for Disease Control and Prevention (see Colford, et al. 1997). The registry contains information submitted by all county health departments within the state. The dataset includes the *P. carinii* diagnosis dates (our index date) for 83735 subjects, which range from 10/87 to 3/97. The county health departments report the date of death for patients to the Office of AIDS. However, this reporting can be delayed, and thus at the time of analysis, the data can contain individuals who have died, but whose deaths have yet to be reported. Ignoring this delay can cause a serious bias if one estimates the survival distribution with the Kaplan-Meier estimator, assuming incorrectly that $T > C$ if $C < V$. As illustrated in the simulation section, this bias should be more severe for estimating the distribution at quantiles with a significant amount of censoring.

First, we estimate the survival distribution for three separate cohorts based on date of diagnosis: 10/87-3/88, 4/88-12/92, and 1/93-3/97. For the most recent cohort, we also include 95% confidence intervals for both the IPCW estimator using (13) and the Kaplan-Meier curve using Greenwood's formula. Next, we estimate the 0.10 quantile for 10 cohorts, one for each quarter from early 1994 through mid 1996. For all these analyses, we estimate the survival distribution both with our IPCW-estimator and the Kaplan-Meier estimator that censors T with C . For the IPCW-estimator, we estimate $G(c)$ with the Kaplan-Meier estimator where C is censored by V .

The results of estimating the survival distribution of the three cohorts is shown in figures 3, 4 and 5. For the earliest cohort, which has few censored subjects, the Kaplan-Meier and the IPCW estimator are nearly identical. However, for the latest cohort (1/93-3/97), the Kaplan-Meier is consistently higher than the IPCW estimator and their confidence intervals do not overlap. This occurs because the Kaplan-Meier estimator distributes the mass of all censored observations to the right of C , and for many of the subjects, their failures could have occurred before the censoring time, i.e. $T < C < V$. This phenomenon can also be seen in figure 5. In this case, the Kaplan-Meier estimator of the 0.10 quantile is consistently higher than our IPCW-estimator. Also, note that the confidence intervals are wider for the IPCW estimator, particularly for later quantiles, because the estimation of the variance accounts for the increased variability due to delay (figure 4). Thus, for this data set the Kaplan-Meier estimate of the survival distribution is significantly greater than our IPCW-estimator, particularly for quantiles with significant censoring.

A literal interpretation of the Kaplan-Meier curves would imply that recent cohorts had a greater increase in survival, perhaps attributed to improvements in treatment. However, the IPCW curves suggest that at least some of this apparent improvement is simply due to bias. For example, this analysis suggests that 14% of the apparent increase in 4 year survival estimated using the Kaplan-Meier estimator is due to the bias caused by delay. However, particularly for recent cohorts, there is a danger that the IPCW estimator is biased as well. For instance, the increase in the IPCW estimate of the 0.10 quantile of the most recent cohort considered in figure 5 could be due to increased survival. However, it could also be due bias in the IPCW estimator which results from a violation of assumption (2). In the next section, we will discuss what happens to the IPCW estimator when this assumption is violated and how we might create an ad hoc fix.

5 Large Delay Times

One possible concern with using our proposed estimator comes from assumption (2), which implies that our estimator will fail if there are subjects for which the reporting time is greater than the support of possible censoring times, say τ . Thus, one must be cautious when interpreting our nonparameteric estimator of survival at times, t , for which there is a probability of having a subject with $T < t$ and $V > \tau$. Note that Tu, et al. (1993) adjust for the bias due to delays by imputing the number of deaths from estimation of the delay time distribution. Whereas our estimator requires assumptions on the censoring distribution (2), the estimator of Tu, et al. (1993), in addition to discretization of the time scale and various parametric assumptions, requires an assumption on the bounding of the reporting delays. Thus, it is difficult to compare our estimator with theirs in the presence of large delays.

First, consider (3) again under the assumption that some subjects can have a V greater than τ . Then,

$$E \left\{ \frac{I(T \leq t)\Delta}{\bar{G}(V)} \right\} = E \left\{ \frac{I(T \leq t)}{\bar{G}(V)} E[\Delta | V] \right\}.$$

Now,

$$E[\Delta | V] = \int I(C > V) dG(C) = I(V < \tau) \bar{G}(V)$$

so the above expectation becomes,

$$E \left\{ \frac{I(T \leq t)}{\bar{G}(V)} I(V < \tau) \bar{G}(V) \right\} = E[I(T \leq t) I(V < \tau)] = P(T \leq t, V < \tau).$$

Note, $F(t)$ can be represented as,

$$F(t) = \frac{P(T \leq t, V < \tau)}{P(V < \tau | T \leq t)},$$

so a corrected estimate of $F(t)$ could be calculated if one had an independent measure of $P(V < \tau | T \leq t)$.

To explore this, we performed simulations for which there was a probability of reporting delays beyond the support of censoring (i.e., for some subjects, one will never know their failure time). The data-generating distributions were the same as in simulation B above with the exception that for each subject, there is a random probability that the reporting time will be essentially infinite. We performed the results for 4 probabilities ($P(V > \tau) = 0.00, 0.05, 0.10, \text{ and } 0.20$) and the results are presented in figure 6, which shows that the bias of the IPCW estimator becomes significant (although less than the Kaplan-Meier) as the probability of long delays increases.

For this data, where the probability infinite reporting delay is independent of T , then one could simply correct the estimate by dividing by $P(V < \tau)$ if one happened to know this probability (in this case, the event of $V > \tau$ is independent of T). When one does not have an estimate of $P(V < \tau | T \leq t)$ and the researcher believes that there is a significant probability of long delays, then at least a sensitivity analysis can be performed on the IPCW estimator, by dividing by the smallest $P(V < \tau | T \leq t)$ one is willing to consider.

For instance, in the AIDS survival example above, consider the analysis of the latest cohort, 1/93-3/97 when the data is analyzed (last data update) on 4/97. In this case, maximum censoring time is the number of days in between 1/93 (earliest diagnosis in the cohort) and the analysis time (4/97), or just under 1600 days ($\tau = 1600$). In this case, one can use historical data to roughly estimate $P(V < \tau | T \leq t)$. Specifically, if one looks at the earliest cohort (last quarter of 1987) which has little censoring (6%), then among the uncensored observations 93% of the observations have a reporting time less than 1600 days if their death was less than 1500 days (the latest quantile we can estimate in this cohort). So, in this case, if the delay distribution of this earliest cohort is any reflection of the cohort of interest, it appears as though our estimator should be relatively unbiased at $t < 1500$. In general, as

progressively later cohorts are examined, the danger increases that (2) is violated for any fixed t . For historical data such as this, at least one can get ranges of possible values of $P(V < \tau | T \leq t)$ so a reasonable sensitivity analysis can be done.

Acknowledgements

This research was supported by a FIRST award from the National Institute of General Medical Sciences, National Institute of Health. We are grateful to Dr. Richard Sun and Jim Creeger of the California State Office of AIDS for providing the data.

References

- P.K. Anderson, O. Borgan, R.D. Gill and N. Keiding, *Statistical Models Based on Counting Processes*, Springer-Verlag, New York, 1993.
- P. Bacchetti “Reporting delays of deaths with AIDS in the United States”, *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* vol. 13 pp. 363–67, 1996.
- P.J. Bickel, A.J. Klaassen, Y. Ritov and J.A. Wellner, *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore, 1993.
- J.M. Colford, M. Sega, F. Tabnak, M. Chen, R. Sun and I. Tager, “Temporal trends and factors associated with survival after *Pneumocystis carinii* Pneumonia in California, 1983–1992”, *American Journal of Epidemiology*, vol. 146 pp. 115–127, 1997.
- R.D. Gill, M.J. van der Laan and J.M. Robins, “Coarsening at Random: Characterizations, Conjectures and Counter-Examples”, *Proceedings of the First Seattle Symposium in Biostatistics*, 1995. D.Y. Lin and T.R. Fleming (editors), Springer Lecture Notes in Statistics, pp. 255–294, 1997.
- D.F. Heitjan and D.B. Rubin, “Ignorability and coarse data”, *Ann. of Statist.* vol. 19 pp.

2244–53, 1991.

P.H. Hu and A.A. Tsiatis, “Estimating the survival function when ascertainment of vital status is subject to delay”, *Biometrika* vol. 83 pp. 371–80, 1996.

A.E. Hubbard, M.J. van der Laan M.J. and J.M. Robins, “Nonparametric locally efficient estimation of the treatment specific survival distribution with right-censored data and covariates in observational studies”, *Statistical Models in Epidemiology: The Environment and Clinical Trials*, E. Halloran and D. Berry (editors), Springer-Verlag, New York, pp. 135–178, 1999.

M. Jacobsen and N. Keiding, “Coarsening at random in general sample spaces and random censoring in continuous time”, *Ann. Statist.* vol. 23 pp. 774–86, 1995.

J.M. Robins and A. Rotnitzky, “Recovery of information and adjustment for dependent censoring using surrogate markers”, *Aids Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz and V.T. Farewell (editors), Birkhäuser, Boston, pp. 297–331, 1992.

J.M. Robins, “Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers”, *Pro. Biopharm. Sec., Am. Stat. Assoc.*, pp. 24–33, 1993.

X. Tu, X. Meng and M. Pagano, “The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data”, *Journal of the American Statistical Association* vol. 88 pp. 26–36, 1993.

M.J. van der Laan and A.E. Hubbard, “Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of the data is delayed”, *Biometrika* vol. 85 pp. 771–83, 1998.

Figure 1: Results of simulations A and B comparing both the Kaplan-Meier and the IPCW estimators to the true survival distribution.

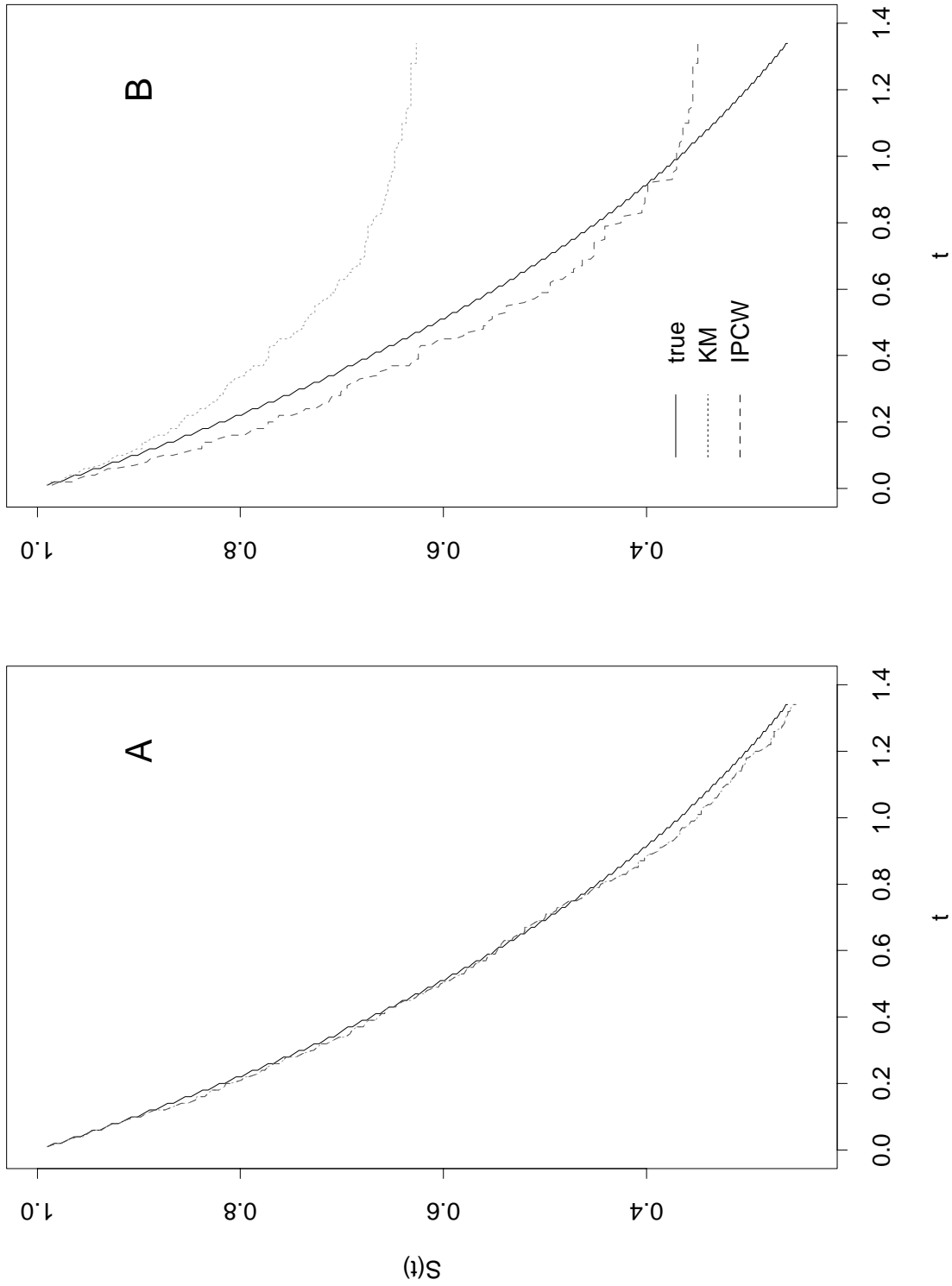


Figure 2: Results of simulation B with 95% pointwise confidence intervals.

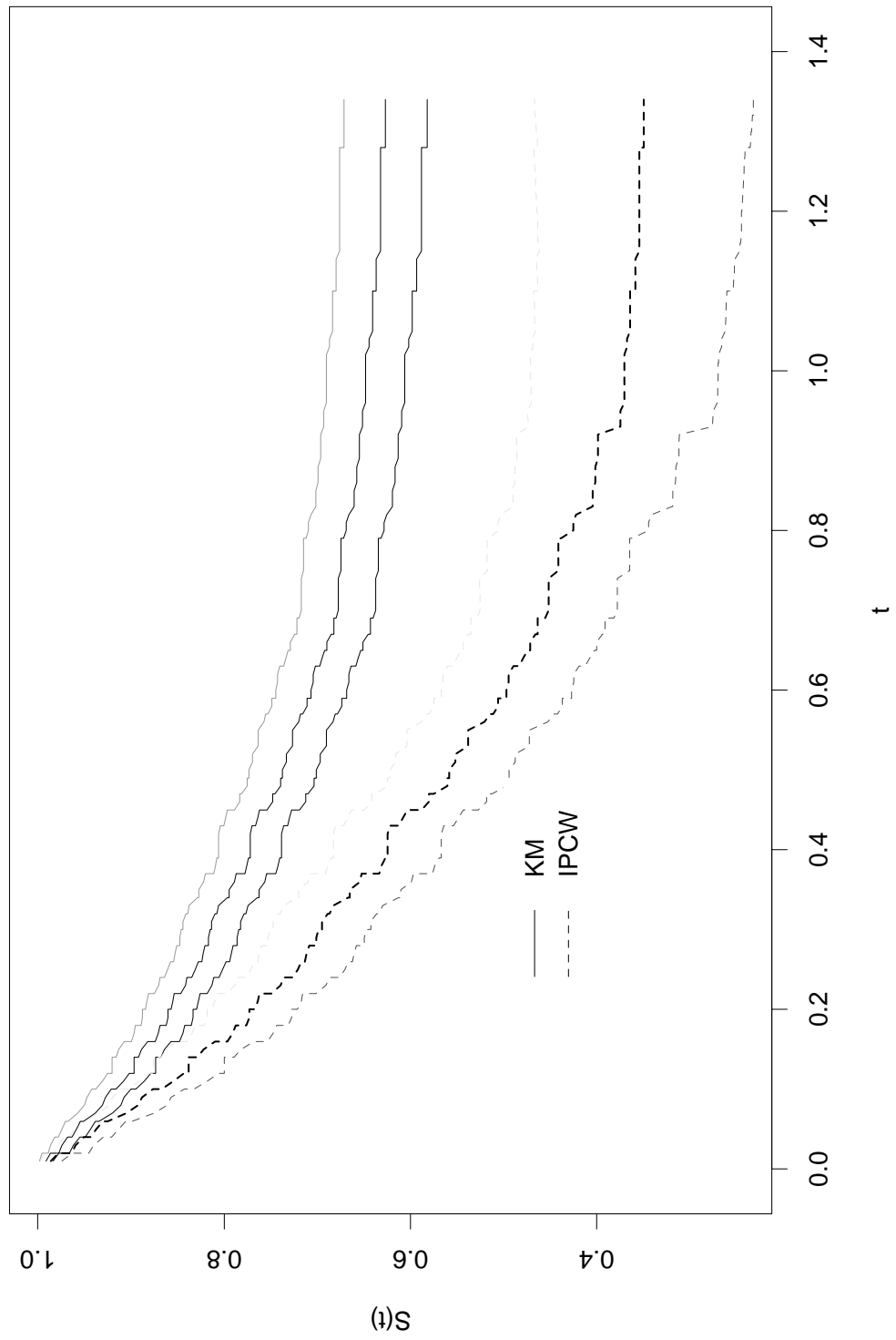


Figure 3: Estimate of survival for three cohorts using both the Kaplan-Meier and the IPCW estimators.

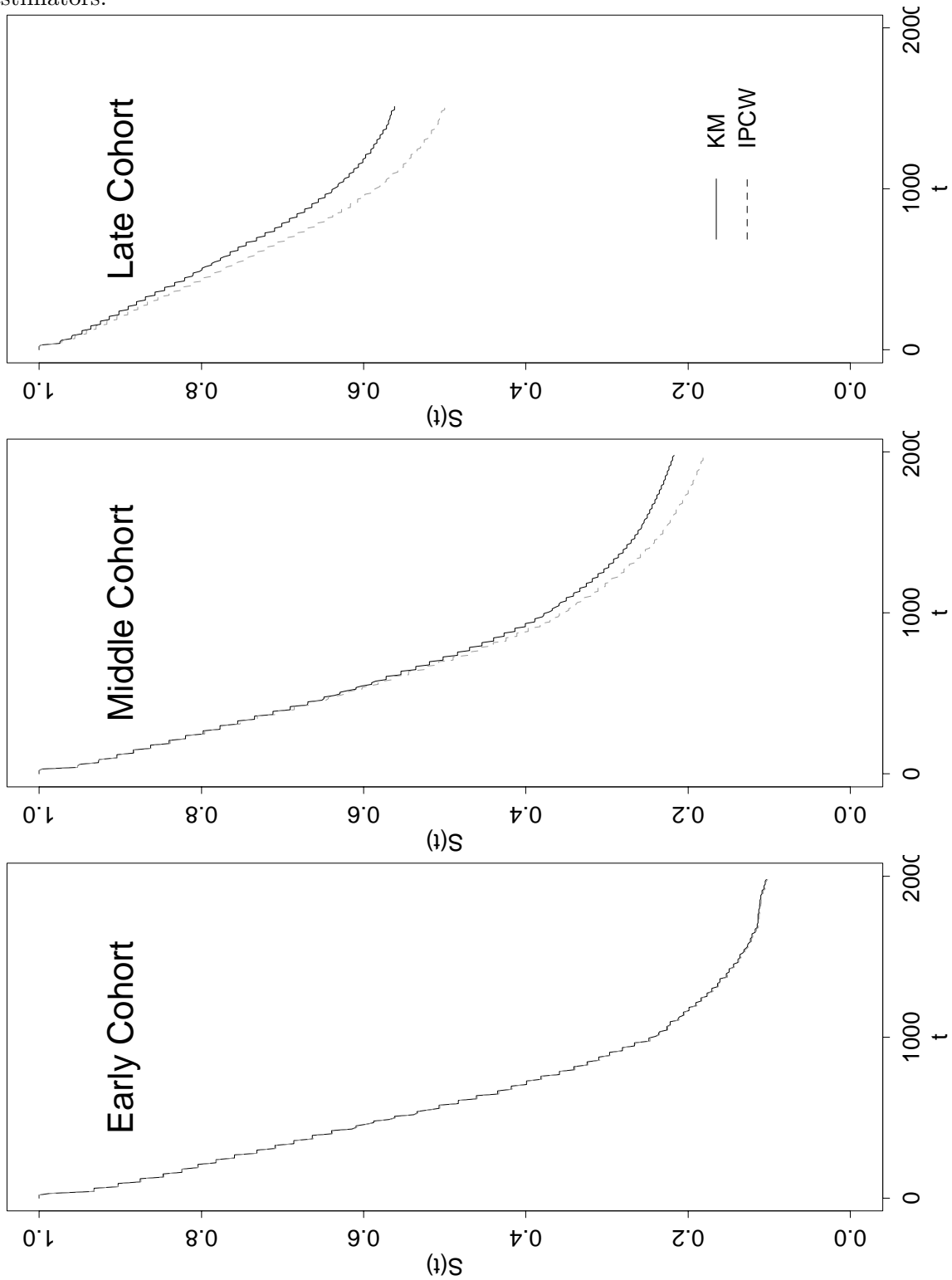


Figure 4: Estimate of survival for 1/93-3/97 cohort and their 95% pointwise confidence intervals using both the Kaplan-Meier and the IPCW estimators.

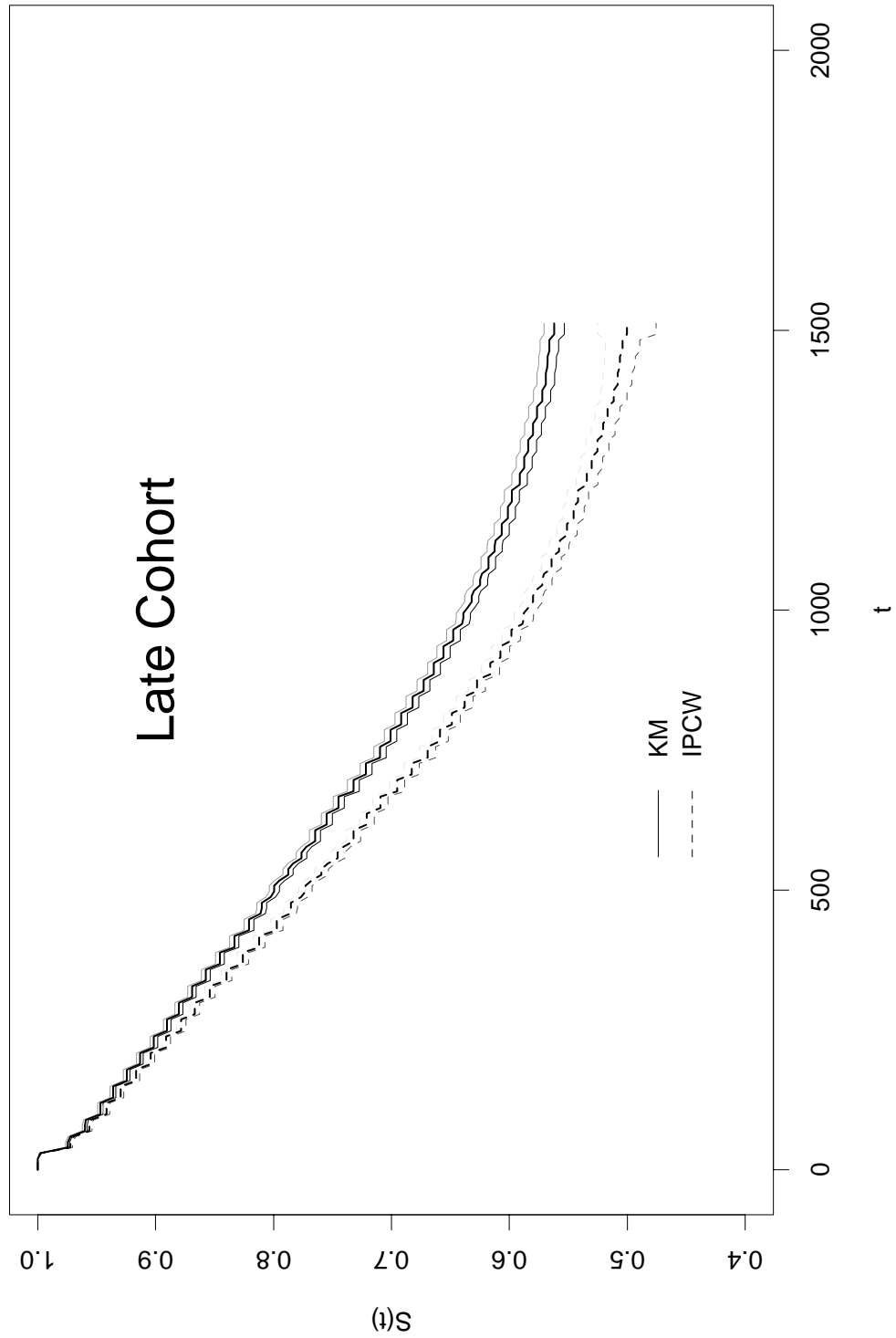


Figure 5: Estimate of 0.10 quantile by cohort defined by quarter of diagnosis using both the Kaplan-Meier and the IPCW estimators.

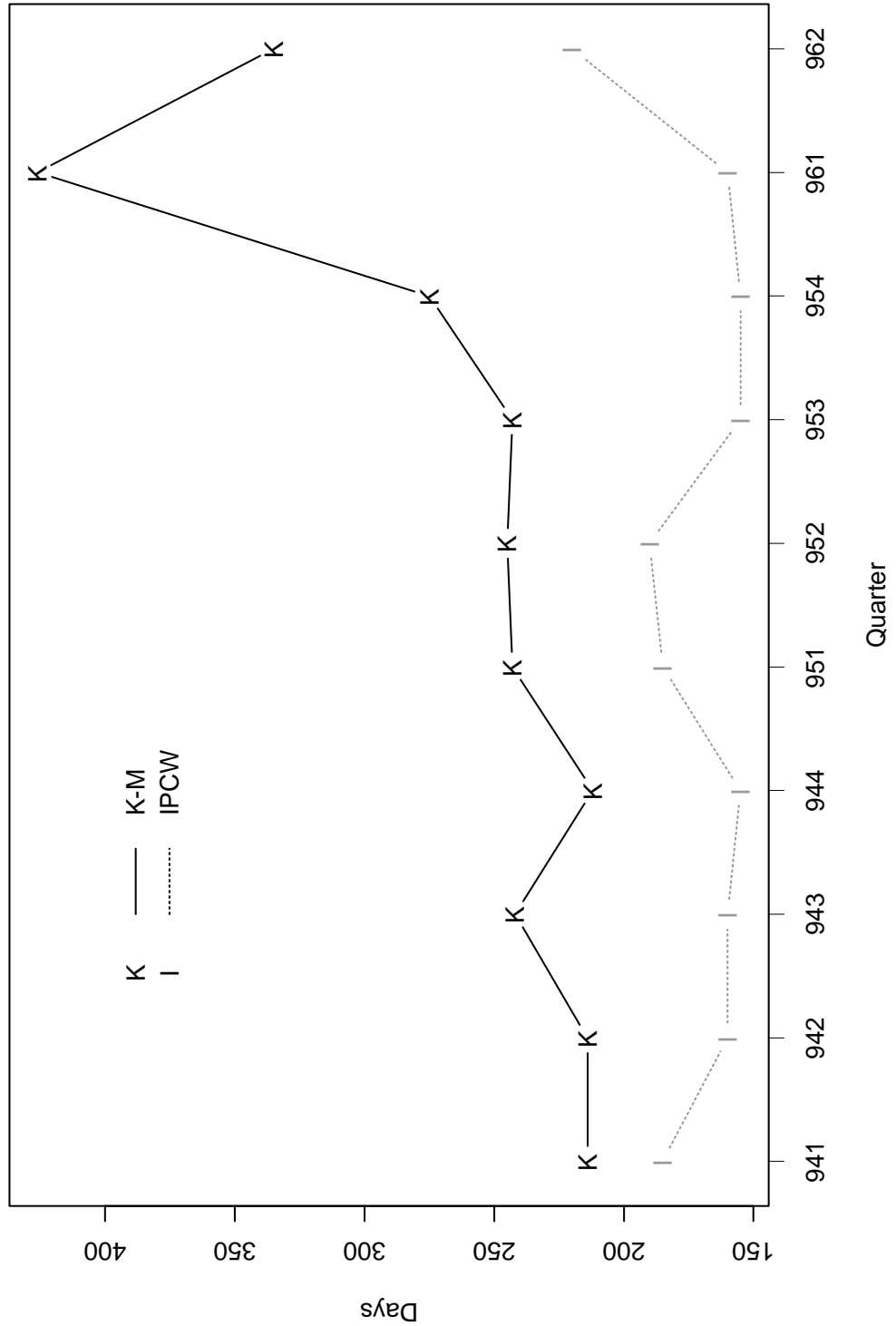


Figure 6: Results of simulations comparing both the Kaplan-Meier and the IPCW estimators to the true survival distribution with increasing probability of infinite reporting delay.

