8-16-2007

INFERENCE FOR SURVIVAL CURVES WITH INFORMATIVELY COARSENED DISCRETE EVENT-TIME DATA: APPLICATION TO ALIVE

Michelle Shardell
Department of Epidemiology and Preventive Medicine, University of Maryland; Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, mshardel@jhsph.edu

Daniel O. Scharfstein
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

David Vlahov
Center for Urban Epidemiologic Studies, New York Academy of Medicine

Noya Galai
Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health

Inference for Survival Curves with Informatively Coarsened Discrete Event-Time Data: Application to ALIVE

Michelle Shardell,[1,*] Daniel O. Scharfstein,[2]
David Vlahov[3], Noya Galai[4]

August 16, 2007

[1]Department of Epidemiology and Preventive Medicine, University of Maryland
660 West Redwood Street Baltimore, Maryland 21201, U.S.A.
[2]Department of Biostatistics, Johns Hopkins University
615 North Wolfe Street Baltimore, Maryland 21205, U.S.A.
[3]Center for Urban Epidemiologic Studies, New York Academy of Medicine
1216 Fifth Avenue New York, New York 10029, U.S.A.
[4]Department of Epidemiology, Johns Hopkins University
615 North Wolfe Street Baltimore, Maryland 21205, U.S.A.
*email: mshardel@jhsph.edu

Abstract

In many prospective studies, including AIDS Link to the Intravenous Experience (ALIVE), researchers are interested in comparing event-time distributions (e.g., for human immunodeficiency virus seroconversion) between a small number of groups (e.g., risk behavior categories). However, these comparisons are complicated by participants missing visits or attending visits off schedule and seroconverting during this absence. Such data are interval-censored, or more generally, coarsened. Most analysis procedures rely on the assumption of non-informative censoring, a special case of coarsening at random that may produce biased results if not valid. Our goal is to perform inference for estimated survival functions

1

across a small number of goups in the presence of informative coarsening. To do so, we propose methods for frequentist and Bayesian inference of ALIVE data utilizing information elicited from ALIVE scientists and an AIDS epidemiology expert about the visit compliance process.

Key Words: Survival Analysis; Informative Censoring; Interval Censoring; Coarsening at Random; Sensitivity Analysis, EM algorithm, MCMC, Gibbs Sampler, Data Augmentation.

1 Introduction

Begun in 1988, AIDS Link to the Intravenous Experience (ALIVE) is an ongoing prospective observational study of risk factors for human immunodeficiency virus (HIV) infection among injection drug users (IDUs) in Baltimore, Maryland. In this study, HIV-negative participants were recruited by community outreach and interviewed upon enrollment regarding drug-related behaviors and other potential HIV risk factors (Vlahov *et al.* 1991, Strathdee *et al.* 2001, Nelson *et al.* 2002). HIV serostatus, a proxy for HIV infection status, was determined by subsequent regularly scheduled laboratory blood tests. For those who attended every visit on schedule, time to seroconversion (years from enrollment) is known, resulting in discrete event-time data. However, ALIVE participants often missed visits or attended visits off schedule, sometimes resulting in seroconversion times only known within a range of years, thus producing interval-censored data. In addition, some seropositive participants never tested positive during the study due to loss to follow up, administrative censoring, or death.

2

AIDS epidemiologists are interested in comparing seroconversion incidence between those who self-reported sharing needles for injecting drugs in the six months prior to enrollment and those who did not. Interval-censored event-time data are usually analyzed by assuming non-informative censoring, a special case of coarsening at random (CAR) (Heitjan and Rubin, 1991; Heitjan, 1993; Gill *et al.*, 1997). However, David Vlahov and Noya Galai, principal investigator and lead statistician of ALIVE, respectively, believe visit compliance may be related to serostatus. That is, data are coarsened not at random (CNAR). Estimation of survival functions with informative coarsening is complicated, as the relationship between censoring and event-time processes is not identified by observed data. Our goal is to extend the frequentist and Bayesian estimation methodology for CNAR data developed in Shardell *et al.* (2006) for application to ALIVE. To do so, we the extend survival curve estimation procedures to address the competing risk of death in ALIVE and propose methods using the survival curve estimates to perform inference regarding the association between sharing needles for injecting drugs and HIV incidence.

When performing inference for survival functions, analyses consist of estimated survival curves and a hypothesis test of equality. In this paper, we propose a class of test statistics utilizing estimates from methods of Shardell *et al.* (2006). Performing inference for several assumed coarsening processes can help assess the sensitivity of scientific conclusions to assumptions. Tests for discrete or grouped continuous interval-censored data have previously been proposed assuming CAR. Sun (1996) and Finkelstein (1986) proposed score tests, and Petroni and Wolfe (1994) proposed a two-sample test for stochastic ordering based on integrated weighted differences (IWD) of survival, all using Turnbull's (1976) estimates. Fay

3

(1996) generalized Finkelstein's (1986) procedure beyond the proportional hazards model. Zhao and Sun (2004) generalized the test proposed by Sun (1996). Akritas (1988) developed a rank-based test. Pan (2000a) suggested multiply imputing all except right-censored event times and performing standard tests for right-censored data. Fang *et al.* (2002) extended Petroni and Wolfe's (1994) test to continuous time. Fay (1999) and Chi (2001) compared these tests' performance. In this paper, we extend the logrank (Mantel, 1966) and a two-sided version of Petroni and Wolfe's (1994) IWD tests to allow informative censoring. We generalize the latter to more than two groups. We also use results from the Bayesian procedure in Shardell *et al.* (2006) to perform inference by proposing a parameter transformation of posterior event-time probabilities, motivated by the logrank test.

The paper is organized as follows. Section 2 describes the data structure, Section 3 provides an overview of CAR, CNAR models, and sensitivity analysis. Section 4 describes the complete-data likelihood and estimation and inference procedures. Section 5 applies the proposed methods to ALIVE. Lastly, section 6 compares and contrasts the methods.

2   Data Structure

Let $T = t$ denote seroconversion during year t, where $E = \{t : t = 1, \ldots, M+1\}$ is the support of T. M denotes last year of follow-up from enrollment, and $T = M+1$ for individuals who did not seroconvert during follow-up. Due to skipped or off-schedule visits, observed data for an individual may be a set of adjacent time periods from E. In particular, observed data are $[L, R] = \{t \in E : L \leq t \leq R\}$. The set $[L, R]$ is a coarsening of T because $T \in [L, R]$. If seroconversion is known

4

to occur in year j, $j = 1, \ldots, M$, then $L = R = j$. If seroconversion did not occur during follow-up, then $L = R = M + 1$, and if knowledge about T is incomplete, then $L < R$. Those with $L < R = M + 1$ are right-censored drop-outs, and those with $L < R < M + 1$ are interval-censored returners.

A complication in ALIVE is the competing risk of death. Those with first missed visit in year l who die in year r either seroconverted in $[L = l, \ R = r]$ or died seronegative. For those censored by death, $[L, \ R]$ has an altered interpretation: R denotes year of death, and possible event times are $\{l, \ldots, r, \ M + 1\}$. Therefore, $T = M + 1$ denotes not seroconverting while at risk. Let $\Delta = \delta$, $\delta \in \{0, 1\}$, indicate whether R is year of death. If $R < M + 1$ and serostatus is unknown at year R due to death, then $\Delta = 1$, otherwise $\Delta = 0$.

Let G denote number of groups. We assume that, for those in group g, $g = 1 \ldots G$, we observe n_g i.i.d. copies of the data. $P_g(\cdot)$ refers to probabilities for those in group g. Where necessary, the subscript i will denote subject-specific data.

3   Coarsening at Random and CNAR Models

In this section we formally define CAR in the context of ALIVE and describe "exponential tilt models" (Barndorff-Nielsen and Cox, 1989) that allow departures from CAR.

3.1   Coarsening at Random

Given $L = l$, $R = r$, and $\Delta = \delta$, let $A(l, r, \delta)$ denote possible values of T, where $A(l, r, 0) = [l, r]$ and $A(l, r, 1) = \{[l, r], \ M + 1\}$. Within group g, CAR means
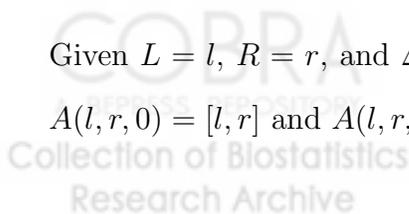
$$P_g(L = l, \ R = r, \ \Delta = \delta \mid T = t) \text{ is constant in } t \in A(l, r, \delta), \qquad (3.1)$$

for all $[l, r] \in E^* = \{[l, r] : l \leq r, \ l, r \in E\}$ and $\delta \in \{0, \ 1\}$. Gill *et al.* (1997) showed Equation (3.1) and

$$P_g(T = t \mid L = l, \ R = r, \ \Delta = \delta) = P_g(T = t \mid T \in A(l, r, \delta)), \qquad (3.2)$$

for all $t \in A(l, r, \delta)$ whatever be $[l, r] \in E^*$, both define CAR.

CAR means, among those in group g, and given vital status at R, the coarsening process provides no information about the seroconversion process beyond knowing the true year of seroconversion is in a set of years. As a result, estimated event-time probabilities for censored individuals only depend on estimated probabilities for years in the set.

3.2 CNAR Models

CAR cannot be identified from observed data, therefore we consider a class of CNAR models indexed by a (possibly group-specific) *censoring bias function*. This function loosens the CAR assumption by allowing elicited expert information to determine whether event probabilities for interval-censored individuals should be made stochastically larger at later times (seroconversions tend to occur late in the censoring interval) or earlier times (seroconversions tend to occur early in the censoring interval) relative to CAR. We exponentially tilt the CAR model for each group g, $g = 1, \ldots, G$:

6
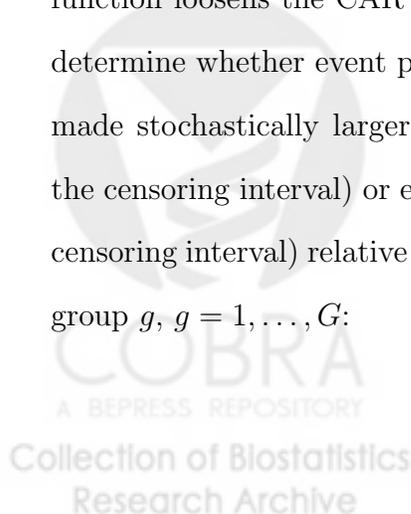
$$P_g(T = t \mid L = l, R = r, \Delta = \delta) = \frac{P_g(T = t \mid T \in A(l, r, \delta)) \exp\{q_g(t, l, r, \delta)\}}{c_g(l, r, \delta \; q_g)},$$

(3.3)

where $c_g(l, r, \delta; q_g) = \sum_{s \in A(l,r,\delta)} P_g(T = s \mid T \in A(l, r, \delta)) \exp\{q_g(s, l, r, \delta)\}$, and $q_g(t, l, r, \delta)$ is a specified censoring bias function of (t, l, r, δ) for those in group g. If $q_g(\cdot)$ does not depend on t, then no tilting is performed, and CAR is assumed for group g. Information about death is only utilized to define possible seroconversion times and in $q(\cdot)$ to allow estimation of the seroconversion process without requiring estimation of the death process.

Using Bayes' rule, Equation (3.3) can be represented as a selection model:

$$\log\left\{\frac{P_g(L = l, \; R = r, \; \Delta = \delta \mid T = t)}{P_g(L = l, \; R = r, \; \Delta = \delta \mid T \in A(l, r, \delta))}\right\} = d_g(l, r, \delta; q_g) + q_g(t, l, r, \delta)$$

(3.4)

for $t \in A(l, r, \delta)$, where $d_g(l, r, \delta; q_g) = -\log\{c_g(l, r, \delta; q_g)\}$. Equation (3.4) implies

$$\log\left\{\frac{P_g(L = l, R = r, \Delta = \delta \mid T = t)}{P_g(L = l, R = r, \Delta = \delta \mid T = t')}\right\} = q_g(t, l, r, \delta) - q_g(t', l, r, \delta),$$
(3.5)

for $t, t' \in A(l, r, \delta)$. From (3.5), we see $q_g(t, l, r, \delta)$ is the difference in log probability of having censoring set $A(l, r, \delta)$ comparing a group g individual with $T = t$ to a group g individual with T equal to some reference value, t_{ref}, such that $q_g(t_{ref}, l, r, \delta) = 0$.

Pattern-mixture models can also be used to interpret $q(t, l, r, \delta)$. From Equation (3.3),

$$\log \left\{ \frac{P_g(T = t \mid L = l, R = r, \Delta = \delta)}{P_g(T = t' \mid L = l, R = r, \Delta = \delta)} \right\} = \tag{3.6}$$

$$\log \left\{ \frac{P_g(T = t \mid T \in A(l, r, \delta))}{P_g(T = t' \mid T \in A(l, r, \delta))} \right\} + q_g(t, l, r, \delta) - q_g(t', l, r, \delta).$$

For those in group g, Equation (3.6) shows $q_g(t, l, r, \delta)$ is the difference in log probability ratios of seroconverting at year t compared to t_{ref}, conditioned on $L = l, R = r$, and $\Delta = \delta$ versus conditioning on $T \in A(l, r, \delta)$ (i.e., versus CAR).

3.2.1 Low-dimensional Parameterization of $q_g(\cdot)$

To facilitate a sensitivity analysis, we parameterize a low-dimensional censoring bias function by a small set of unidentified *censoring bias parameters* to capture key features of ALIVE. The function is indexed by parameters differentiating between those who are interval censored, right-censored alive, and censored by death. We allow the censoring mechanism to differ between needle sharers and non-sharers. Let $\phi = \{\phi_g : g = 1, 2\}$ denote group-specific censoring bias parameters, where $g = 1$ ($g = 2$) denotes non-sharers (needle sharers). The proposed censoring bias function is

$$q_g(\phi, t, l, r, \delta) = \frac{9}{4} \phi_{g1} I(r < M + 1)(1 - \delta) \frac{(t - l)}{(M - 1)} + \phi_{g2} I(r = M + 1) \frac{(t - l)}{M}$$

$$+ \phi_{g3} I(r < M + 1)(\delta) \frac{(t - l)}{M}, \ g = 1, 2, \tag{3.7}$$

where $\phi_g = \{\phi_{g1}, \phi_{g2}, \phi_{g3}\}$. Using Bayes' rule as in Section 3,

8

- $\exp\{\phi_{g1}\}$ is the needle sharing-specific probability ratio of having interval [1 year, 5 years] comparing those who seroconvert during the year five to those who seroconvert during the first year.

- $\exp\{\phi_{g2}\}$ is the needle sharing-specific probability ratio of dropping out after baseline comparing those who do not seroconvert within ten years to those who seroconvert during the first year, among those who remain alive throughout the study.

- $\exp\{\phi_{g3}\}$ is the needle sharing-specific probability ratio of dropping out after baseline comparing those who do not seroconvert while alive to those who seroconvert during the first year, among those who die during the study.

The factor $\frac{9}{4}$ accounts for the ten year follow-up, but investigators were more comfortable stating beliefs for a five-year interval than for a ten-year interval. When $\exp\{\phi_{g1}\} > 1 \ (< 1)$, returners are assumed to be more (less) likely to seroconvert late than seroconvert early. When $\exp\{\phi_{g2}\} > 1 \ (< 1)$, drop-outs who remain alive are assumed more (less) likely to seroconvert late or not at all than seroconvert early. When $\exp\{\phi_{g3}\} > 1 \ (< 1)$, drop-outs who die with unknown serostatus are assumed more (less) likely to seroconvert late or not at all than seroconvert early. Using the pattern-mixture approach, $\exp\{\phi_{gh}\} > 1$ $(\exp\{\phi_{gh}\} < 1)$ means those with needle-sharing status g, $g = 1, 2$, and censoring pattern h, $h = 1, 2, 3$, seroconvert stochastically later (earlier) than expected assuming CAR.

9

3.3 Frequentist Inference

Frequentist estimation is performed using the EM algorithm. The complete-data likelihood is $\mathrm{L}(\boldsymbol{p}) = \prod_{g=1}^{G} \prod_{i=1}^{n_g} \prod_{j=1}^{M+1} p_{gj}^{I_{igj}}$. Initial estimates of $\boldsymbol{p}_g$ are used to evaluate the expected value of the complete-data log likelihood, given observed data (E-step). The E-step at iteration s is

$$Q(\boldsymbol{p}; \boldsymbol{p}^{(s-1)}) = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{j=1}^{M+1} I_{igj}^{(s-1)} \log(p_{gj}) - \sum_{g=1}^{G} [\lambda_g (\sum_{j=1}^{M+1} p_{gj} - 1)],$$

where λ_g are Lagrange multipliers, and

$$I_{igj}^{(s-1)} = \frac{\omega_{igj} p_{gj}^{(s-1)} \exp\{q_g(j, l, r, \delta)\}}{\sum_{k=1}^{M+1} \omega_{igk} p_{gk}^{(s-1)} c_{ig}^{(s-1)}(l, r, \delta; q_g)}.$$

$Q(\boldsymbol{p}; \boldsymbol{p}^{(s-1)})$ is maximized (M-step) to obtain updated estimates of $\boldsymbol{p}_g$. The M-step results in a reweighted version of Turnbull's (1976) self-consistency equation for each group g, $g = 1, \ldots, G$: $p_{gj}^{(s)} = \frac{1}{n_g} \sum_{i=1}^{n_g} \frac{\omega_{igj} p_{gj}^{(s-1)} \exp\{q_g(j,l,r,\delta)\}}{\sum_{k=1}^{M+1} \omega_{igk} p_{gk}^{(s-1)} \exp\{q_g(k,l,r,\delta)\}}$. When $q_g = 0$ (CAR assumed), our estimator simplifies to that of Turnbull (1976). Standard errors for probabilities in each group are estimated using Louis's (1982) method. Probability estimates are used to estimate $S_g(\cdot)$, the survivor function for group g. Once $\hat{\boldsymbol{p}}$ and standard errors are obtained, statistics can be derived for testing the null hypothesis $H_0 : S_1(\cdot) = \ldots = S_G(\cdot) = S(\cdot)$ using the delta method, including logrank (LR) (Mantel, 1966) and IWD (Petroni and Wolfe, 1994) tests.

Let $LR = (LR_1, \ldots, LR_G)^t$ be a vector of length G with gth component $LR_g = \sum_{j=1}^{M} \left(d_{jg} - n_{jg} \frac{d_j}{n_j} \right)$, where $d_{jg} = n_g \hat{p}_{jg}$ is the estimated number of seroconverts in group g during year j, $n_{jg} = n_g \sum_{k=j}^{M+1} \hat{p}_{jg}$ is the estimated number at risk in group g during year j, $d_j = \sum_{g=1}^{G} d_{jg}$, and $n_j = \sum_{g=1}^{G} n_{jg}$. The variance of LR, Σ_{LR}, is a $G \times G$ matrix estimated by $\hat{\Sigma}_{LR}$. The logrank test statistic

10

is $\chi^2_{LR} = LR^t \hat{\Sigma}^-_{LR} LR$, where X^- denotes generalized inverse of square matrix X. Under the null hypothesis, χ^2_{LR} has a χ^2 distribution with $G - 1$ degrees of freedom.

Petroni and Wolfe's (1994) IWD test involved a one-sided hypothesis for the special case of $G = 2$. This test can be generalized to a two-sided test. The two-sample test with weight $w(\cdot)$, estimated by $\hat{w}(\cdot)$, has numerator $IWD = \sum_{j=1}^{M} \hat{w}(j) \left[\hat{S}_1(j) - \hat{S}_2(j) \right]$, where $\hat{S}_g(j) = \sum_{k=j+1}^{M+1} \hat{p}_{gk}$. Let σ^2_{IWD} denote the variance of IWD, estimated by $\hat{\sigma}^2_{IWD}$. The test statistic, $Z_{obs} = IWD / \hat{\sigma}_{IWD}$, can be compared to a standard normal distribution, following large-sample theory presented in Petroni and Wolfe (1994). When $G \geq 2$, the test can be modified by comparing $\hat{S}_g(j)$, $g = 1, \ldots, G$, to the estimated overall survivor function: $\hat{S}(j) = \sum_{g=1}^{G} \frac{n_g \hat{S}_g(j)}{n}$. $IWD = (IWD_1, \ldots, IWD_G)^t$ is a vector of length G with gth component $IWD_g = \sum_{j=1}^{M} \hat{w}(j) \left[\hat{S}_g(j) - \hat{S}(j) \right]$. The variance of IWD is a $G \times G$ matrix, Σ_{IWD}, estimated by $\hat{\Sigma}_{IWD}$. The test statistic, $\chi^2_{IWD} = IWD^t \hat{\Sigma}^-_{IWD} IWD$, is distributed χ^2 with $G - 1$ degrees of freedom under the null hypothesis.

Simulation studies (Appendix 1) showed IWD and LR tests perform well and are accurately sized.

3.4 Bayesian Inference

We assume a Dirichlet distribution for the probability of seroconversion for each year. Let $B_g = \{b_{g1}, \cdots, b_{g(M+1)}\}$ be a base measure defined on E for those in group g, the prior mean of $\boldsymbol{p}_g$. A precision parameter, α^*, describes concentration of the distribution around B_g, where elements of B_g sum to 1. Let $\alpha_{gj} = \alpha^* b_{gj}$,

11

for $j = 1, \ldots, M + 1$. The Dirichlet density is given by:

$$f(\boldsymbol{p}_g) = \frac{\Gamma(\alpha_{g1} + \ldots + \alpha_{g(M+1)})}{\Gamma(\alpha_{g1}) \ldots \Gamma(\alpha_{g(M+1)})} \prod_{j=1}^{M+1} p_{gj}^{\alpha_{gj}-1}, \tag{3.8}$$

where $p_{g1}, \cdots, p_{g(M+1)} \geq 0$; $\sum_{k=1}^{M+1} p_{gk} = 1$; the $\boldsymbol{\alpha}_g = \{\alpha_{g1}, \ldots, \alpha_{g(M+1)}\}$ are all positive; and α_{gj}'s are interpreted as 'prior counts' of seroconverts during year j in group g. We assume the censoring bias function for group g, q_g, is indexed by a vector of (possibly group-specific) censoring bias parameters, $\boldsymbol{\phi}$. Data are incomplete, hence conjugate analyses like those in Calle and Gomez (2001) cannot be performed. Therefore, we propose analysis via Markov Chain Monte Carlo (MCMC) using the Gibbs sampler (Geman and Geman, 1984) with data augmentation as in Tanner and Wong (1987) and a Metropolis-Hastings step (Hastings, 1970). The detailed algorithm is in Appendix 2.

Simulated $\boldsymbol{p}$'s can be transformed into a one-dimensional quantity summarizing the difference between G event-time distributions. The proposed quantity is motivated by the logrank test. Let $LR(\boldsymbol{p}^{(s)})$ denote the posterior logrank transformation at iteration s, a vector of length G with gth component $LR_g(\boldsymbol{p}^{(s)}) = \sum_{j=1}^{M} \left(d_{jg}^{(s)} - n_{jg}^{(s)} \frac{d_j^{(s)}}{n_j^{(s)}} \right)$, where $d_{jg}^{(s)} = n_g p_{jg}^{(s)}$, $n_{jg}^{(s)} = n_g \sum_{k=j}^{M+1} p_{jg}^{(s)}$, $d_j^{(s)} = \sum_{g=1}^{G} d_{jg}^{(s)}$, and $n_j^{(s)} = \sum_{g=1}^{G} n_{jg}^{(s)}$. Let $\Sigma_{LR}(\boldsymbol{p}^{(s)})$ be a $G \times G$ matrix motivated by the variance of the logrank test numerator when the null hypothesis is true: $\Sigma_{LR}(\boldsymbol{p}^{(s)}) = \sum_{j=1}^{M} \frac{d_j^{(s)}(n_j^{(s)} - d_j^{(s)})(n_j^{(s)} n_{jg'}^{(s)} I(g=g') - n_{jg}^{(s)} n_{jg'}^{(s)})}{(n_j^{(s)})^2 (n_j^{(s)} - 1)}$. The transformation is complete by calculating $\chi^2(\boldsymbol{p}^{(s)}) = LR(\boldsymbol{p}^{(s)})^t \Sigma_{LR}^-(\boldsymbol{p}^{(s)}) LR(\boldsymbol{p}^{(s)})$. Let N_{sim} denote the number of Gibbs sampler iterations. The median of the parameter transformation under the null hypothesis is approximately $\mu_G = G - 1 - \frac{2}{3} + \frac{4}{27(G-1)} - \frac{8}{729(G-1)^2}$. The vector of logrank parameter transformations is denoted

12

by $\chi^2(\boldsymbol{p}) = \{\chi^2(\boldsymbol{p}^{(1)}), \cdots \chi^2(\boldsymbol{p}^{(N_{sim})})\}$. A posterior tail probability can be calculated by $2P\left(\chi^2(\boldsymbol{p}) \leq \mu_G \mid \boldsymbol{\omega}\right)$. In addition, $\chi^2(\boldsymbol{p})$ can be plotted with a χ^2_{G-1} kernel, the distribution of the logrank test when all G event-time distributions are equal. When $G = 2$, a transformation can be calculated by $Z(\boldsymbol{p}^{(s)}) = \frac{\mathrm{LR}_2(\boldsymbol{p}^{(s)})}{\sigma(\boldsymbol{p}^{(s)})}$, where $\sigma(\boldsymbol{p}^{(s)})$ is the standard deviation of $LR_2(\boldsymbol{p}^{(s)})$. Let $Z(\boldsymbol{p}) = \{Z(\boldsymbol{p}^{(1)}), \cdots Z(\boldsymbol{p}^{(N_{sim})})\}$. $Z(\boldsymbol{p})$ can be plotted with a standard normal kernel, and the tail probability can be calculated as $2\left[\min\{P\left(Z(\boldsymbol{p}) \geq 0 \mid \boldsymbol{\omega}\right), P\left(Z(\boldsymbol{p}) \leq 0 \mid \boldsymbol{\omega}\right)\}\right]$. Posterior probabilities are interpreted differently from frequentist p-values. Instead of calculating the tail probability of a test statistic under H_0 at the observed value, we calculate the posterior tail probability of a parameter transformation at its expected value when H_0 is true.

4 ALIVE Data Analysis

We apply our proposed methodology to ALIVE to compare the ten-year incidence of seroconversion between those who self-reported needle sharing at enrollment and those who did not ($G = 2$). Censoring due to missed visits is thought to be informative and may depend on self-resported needle-sharing status. Serostatus was determined by enzyme-linked immunosorbent assay (ELISA). Those who were repeatedly reactive were confirmed by Western blot (WB). Estimated sensitivity and specificity are over 99% for ELISA combined with WB (Chou *et. al.*, 2005), therefore issues regarding misclassified serostatus will not be addressed.

ALIVE consists of 2205 participants with complete needle-sharing information. At baseline, 1527 participants reported sharing needles, while the remaining 678 did not. Among those who reported sharing needles, 12%, 74%, 9%, and 4% were censored by death, right-censored by drop-out or end of study, interval censored,

and exactly observed, respectively. Percentages among those who reported not sharing needles were 11%, 77%, 8%, and 4%. Among needle sharers, 242 (16%) died during follow-up: 52 (21%) died after seroconverting while the remaining 190 (79%) died with unknown serostatus. Among those who reported not sharing needles, 100 (15%) died during the study: 27 (27%) died after seroconverting while the remaining 73 (73%) died with unknown serostatus. The relationship between needle-sharing and seroconversion may be biased due to differential death rates between groups. However, logrank test results (p-value = 0.52) do not support this hypothesis. Those who died during the study may have a different assumed relationship between visit and seroconversion processes than those who remained alive at the end of the study. Death as a primary endpoint was not addressed in the sensitivity analysis.

4.1 Elicitation and Sensitivity Analysis

To elicit values of ϕ in Equation (3.7), Drs. Galai and Vlahov were separately shown Figure 1a and were asked, "Among those who self-reported needle sharing at baseline, who is more likely to test negative for HIV at baseline, miss visits, then return during the fifth year and test positive: one who seroconverted during the first year or one who seroconverted during the fifth year? How many times more likely?" Using Figure 1b, they were asked, "Among those who self-report needle sharing at baseline and who remained alive throughout the study, who is more likely to test negative for HIV at baseline, then drop out: one who seroconverted during the first year or one who did not seroconvert within ten years? How many times more likely?" Lastly, using Figure 1c, they were asked, "Among those who self-reported needle sharing at baseline, who is more likely to test negative

14

for HIV at baseline, then drop out and die with unknown serostatus: one who seroconverted during the first year, or one who did not seroconvert while at risk? How many times more likely?" Questions were repeated for self-reported non-sharers.

Elicited ϕ varied between the two experts. After reaching a consensus about ranges of plausible values, the experts believe, among needle sharers, those who seroconvert during the fifth year are 1.75 times less to 2.75 times more likely to be censored into interval (0 years, 5 years] than those who seroconvert during the first year. The range for non-sharers is 1.15 times less to 2.50 times more likely. The experts expressed uncertainty about the direction of this relationship because those who seroconvert earlier may either behave erratically and miss visits, but return when their health diminishes, or may acknowledge their high-risk status and feel motivated to participate in the study, compared to those who seroconvert later. For those who remained alive after ten years, the experts believe, among needle sharers, those who did not seroconvert within ten years are 1.50 to 3.00 times more likely to drop out after baseline than those who seroconvert within one year. Among non-sharers, the range was 1.75 to 2.50 times more likely. Among those who die within ten years from baseline, the experts believe those who do not seroconvert while at risk are 2.00 to 2.50 more likely to drop out than those who seroconvert within the first year with the same baseline needle-sharing status. The experts anticipate those who seroconvert early would eventually return to the study as their condition worsens, compared to those who do not seroconvert. Those who die with unknown serostatus are likely to die from other reasons, such as drug overdose and homicide.
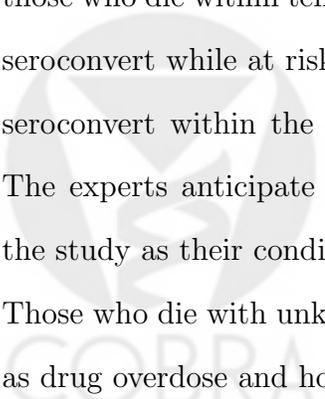
Ranges of elicited values are sufficient for performing a sensitivity analysis

under several fixed assumptions, but additional elicited information is needed to perform a Bayesian analysis mixing over assumptions. Prior beliefs about the seroconversion time distribution were elicited from Dr. Samuel Friedman, an expert on AIDS epidemiology among injection drug users. Dr. Friedman was interviewed about his expected seroconversion time distribution and the weight of his expert opinion relative to ALIVE data. An expert not affiliated with the ALIVE study was purposely chosen, as we are interested in opinion *prior to* ALIVE. HIV incidence depends on seroprevalence in the population (Friedman *et al.* 1995), and, given Baltimore's high HIV seroprevalence of approximately 24% among IDUs in 1988 (Vlahov *et al.* 1991) and HIV prevention efforts (Wiebel and Altman, 1988), Dr. Friedman's prior belief is incidence would decline over time, where $\approx 65\%$ would remain seronegative after ten years. However, he believes his prior opinion should be weighted 10% of the final results (ALIVE data weighted 90%). Prior information about seroconversion probabilities was not specific to needle-sharing status to reflect the "null" belief of equal seroconversion-time distributions.

prior information about the distribution of ϕ was collected from Dr. Vlahov. For each needle-sharing and censoring combination, unimodal histograms from several distributions with various modes and variances were displayed; each a realization from a beta distribution, centered and scaled to reflect elicited ranges of $\exp\{\phi\}$, thus we graphically elicited a prior mode and variance for each censoring bias parameter. Dr. Vlahov believes the distribution of

- $\exp\{\phi_{11}\}$ ($\exp\{\phi_{21}\}$) is right skewed with mode 1.0 (0.80),

- $\exp\{\phi_{12}\}$ ($\exp\{\phi_{22}\}$) is left skewed with mode 2.25 (2.5),

- and $\exp\{\phi_{13}\}$ ($\exp\{\phi_{23}\}$) is flat and left skewed with mode 2.5 (2.5).

16

In addition, Dr. Vlahov believes the distribution for $\exp\{\phi_{22}\}$ is flatter than that for $\exp\{\phi_{12}\}$, reflecting greater uncertainty about behavior of alive needle-sharing drop-outs compared to their non-sharing counterparts. He also indicated these beliefs are correlated. Using scatter plots like those in Figure 2 to elicit this information, Dr. Vlahov's prior variance-covariance matrix involves positive correlations for several groups. The rationale for high correlations within censoring group is believed similarity between groups regarding other HIV risk factors (i.e., sexual behavior) and dishonest needle-sharing reporting. High correlations between drop-out groups reflect that many of those who died with unknown serostatus may not have returned to the study even if they stayed alive during follow-up. Beliefs about interval-censored groups were uncorrelated with drop-out groups, because motivations for visit compliance behavior may differ.

4.2 Frequentist Analysis

Sufficient conditions for unique estimation (Shardell *et al.*, 2006) were met. Frequentist analyses were performed assuming CAR and combinations of minimum and maximum elicited values of $\boldsymbol{\phi}$. For each combination, needle-sharing specific seroconversion probabilities were estimated, and logrank and IWD tests were performed with weights $w = 1$, and w^*, where $w^*(j) \equiv \frac{\prod_{g=1}^{G} \hat{K}_g(j)}{\sum_{g=1}^{G} (n_g/n)\hat{K}_g(j)}$, and $\hat{K}_g(j)$ is the proportion of group g participants with known serostatus in year j.

Table 1 shows estimated needle-sharing specific one-year, five-year, and ten-year seronegative probabilities and 95% confidence intervals (using the complementary log-log transformation) for three values of $\exp\{\boldsymbol{\phi}\}$: CAR ($\boldsymbol{\phi} = 0$), $\{\max(\boldsymbol{\phi}_2),\min(\boldsymbol{\phi}_1)\}$, and $\{\min(\boldsymbol{\phi}_2),\max(\boldsymbol{\phi}_1)\}$. When CAR is assumed, estimated probabilities are similar across groups, corroborated by large p-values for IWD

17

($w = 1$, p-value=0.857; $w = w^*$, p-value=0.866) and logrank (p-value=0.684) tests. The minimum p-values (w=1, p-value=0.170; $w = w^*$, p-value=0.176; LR, p-value=0.109) were produced when needle sharers are assumed to seroconvert stochastically early in their observed sets (minimum ϕ_2) and non-sharers are assumed to seroconvert stochastically late in their observed sets (maximum ϕ_1) according to elicited ranges for ϕ. Estimated probabilities of remaining seronegative for non-sharers are higher than those for needle-sharers for this assumption. Similarly, when the opposite assumption is made (minimum ϕ_1, maximum ϕ_2), estimated seronegative probabilities for non-sharers were lower than those for needle sharers ($w = 1$, p-value=0.528; $w = w^*$, p-value=0.525; LR, p-value=0.675). Other combinations of ϕ_1 and ϕ_2 were also explored (data not shown). Test results were most sensitive to values of ϕ_{22}, the parameter for needle-sharing drop-outs who remain alive, because this group is the largest needle-sharing status-by-censoring type category in ALIVE, and the experts were most uncertain about them.

4.3   Bayesian Analysis

Dr. Friedman's prior beliefs about seroconversion time were converted into Dirichlet hyperparameters. Dr. Vlahov's prior beliefs about censoring bias parameters were converted into multivariate normal hyperparameters. First, $\exp\{\phi\}$ were centered and scaled to have range $[0, 1]$: $\exp\{\phi_{gh}\}_{scaled} = \frac{\exp\{\phi_{gh}\} - \exp\{\min(\phi_{gh})\}}{\exp\{\max(\phi_{gh})\} - \exp\{\min(\phi_{gh})\}}$, where $g = 1$, 2, $h = 1$, 2, 3. These $\exp\{\phi\}_{scaled}$ were assumed to have marginal beta distributions, inducing a mean and variance used to make a normal approximation to the beta density. The normal approximation for $\exp\{\phi\}$ can be found by transforming $\exp\{\phi\}_{scaled}$ back to the elicited range (Table 2). Elicited correlation coefficients then induced an approximate multivariate normal joint

18

distribution for $\exp\{\boldsymbol{\phi}\}$. However, in order to preserve elicited ranges for sampled $\exp\{\boldsymbol{\phi}\}$, $\exp\{\boldsymbol{\phi}\}$ were centered and scaled to have range $[0, 1]$, then the logit transformation was performed, and these transformed parameters were simulated from their multivariate normal distribution.

Before modeling random $\boldsymbol{\phi}$, Bayesian analysis was performed for fixed $\boldsymbol{\phi}$. We examine CAR and the same two extreme specifications discussed in Section 4.2: 1) (non) needle-sharers seroconverting stochastically late (early) and 2) (non) needle-sharers seroconverting stochastically early (late). The Gibbs sampler was run for 500 burn-in and 5000 additional iterations. For this and all subsequent analyses, the diagnostic scheme from Cowles and Carlin (1996) was used. Needle-sharing specific mean posterior one-year, five-year, and ten-year seronegative probabilities (95% credible intervals) are shown in Table 3. The first three columns are Bayesian analogs of frequentist results shown in Table 1. When CAR is assumed, the mean posterior logrank transformation $\{Z(\boldsymbol{p})\}$ (95% credible interval) is 0.407 (-1.935, 2.740) with a tail probability of 0.736. For the first {second} specification, the mean posterior logrank transformation (95% credible interval) is -0.566 (-2.972, 1.618) {1.696 (-0.591, 3.828)} with a tail probability of 0.629 {0.128}. Thus, zero is a plausible value for the mean of $\{Z(\boldsymbol{p})\}$ in all three specifications. Mean posterior seronegative probabilities are lower than estimated probabilities from the EM algorithm with equal $\boldsymbol{\phi}$, shown in Table 1. This result is due to shrinkage to the prior, which suggested more accelerated seroconversion than estimates from data alone. Also, credible intervals are slightly more narrow than analogous confidence intervals, especially for the ten-year seronegative probability, due to additional information from the prior and many drop outs.

19

Next, fully Bayesian analysis was performed, averaging over the posterior distribution of ϕ. The Gibbs sampler was burned in for 1000 iterations and run for 10000 more. Metropolis-Hastings acceptance was 64% and 86% for needle sharers and non-sharers, respectively. Prior and posterior correlations for $\exp\{\phi\}$ are shown in Figure 3. Prior and posterior densities for one-year, five-year, and ten-year probabilities of being seronegative are reported in the first row of Figure 4. Posterior densities are much tighter than priors, due to small weight given to elicited information relative to the data. Mean posterior one-year, five-year, and ten-year needle-sharing specific seronegative probabilities (95% credible intervals) for needle sharers are shown in the last column of Table 5. Seronegative probabilities are between those obtained using extreme elicited censoring bias parameter values. Box plots for prior and posterior distributions for censoring bias parameters are shown in the second row of Figure 4. Marginal posterior distributions are almost identical to the priors, as data provide no information about these parameters. Posterior means (95% credible intervals) of $\exp\{\phi_2\}$ (needle-sharers) are 1.002 (0.677, 1.564), 2.454 (1.939, 2.849), and 2.318 (2.102, 2.468). Posterior means (95% credible intervals) of $\exp\{\phi_1\}$ (non-sharers) are 1.245 (0.965, 1.718), 2.231 (1.997, 2.408), and 2.317 (2.102, 2.469). The posterior mean (95% credible interval) of $Z(\boldsymbol{p})$ is 0.357 (-1.900, 2.494), with tail probability 0.738. Figure 5 shows the posterior logrank parameter transformation differs little from the standard normal kernel, suggesting seroconversion distributions do not differ across needle-sharing status. Posterior mean survival probabilities and tail probability corroborate this conclusion.

20

5 Discussion

Two different procedures for comparing survival curves for informatively coarsened data were developed in this paper and applied to ALIVE. The CAR-based analysis of ALIVE data suggests baseline needle sharing is not significantly associated with time to seroconversion. These results hold for both Bayesian and frequentist sensitivity analysis procedures, and conclusions are robust to elicited assumptions about the visit-compliance process. CAR-based and random-ϕ posterior survival probabilities varied slightly, but qualitative conclusions were identical.

The proposed methods are beneficial in that they enable statisticians and scientists to discuss assumptions about scientific questions and standard statistical procedures. They are more flexible and more honestly represent knowledge about coarsening than methods that solely rely on CAR or any one alternative assumption. Additionally, analysis results can be displayed like those from CAR-based analyses. These methods can be used to design studies, allowing the statistician to build various scientists' assumptions about coarsening into sample size calculations when frequentist procedures will be used, where information from past studies can serve as auxiliary information in planning a subsequent study.

However, these methods have limitations. In particular, results may be sensitive to distributional assumptions. For example, the correlation structure of Dirichlet priors does not take advantage of time ordering of visits. Also, the proposed methods are limited to small numbers of groups.

The proposed methods are not meant to replace objective statistical procedures with subjective ones. When data are coarsened, additional assumptions are required to estimate parameters of interest. Therefore, the best that can be

21

accomplished is a sensitivity analysis based on unidentifiable subjective assumptions (Kadane, 1993). Although scientists with discordant opinions may derive different conclusions from the same analysis, it is our hope such differences would facilitate productive discussions within the scientific community.

Appendix 1: Simulation Study

Simulations were performed for two-sample and G-sample logrank and IWD tests with $G = 3$, allowing left censoring, and no competing risks. Event times were simulated from a multinomial distribution using the continuation ratio logistic model with $M = 4$. Let $\rho_{ij} = P(T_i = j \mid T_i \geq j, \boldsymbol{Z}_i)$ for $j = 0, ..., M + 1$. The continuation ratio model for the three-sample simulation is $\log\left(\frac{\rho_{ij}}{1 - \rho_{ij}}\right) = \theta_j + \boldsymbol{\beta}\boldsymbol{Z}_i^t$, $j = 0, \ldots, M$, where $\boldsymbol{\beta} = \{\beta_1, \beta_2\}$ and $\boldsymbol{Z}_i = \{Z_{i1}, Z_{i2}\}$, where $Z_1 = I(g = 2)$ and $Z_2 = I(g = 3)$. Replacing $\boldsymbol{\beta}\boldsymbol{Z}_i^t$ with $\beta_1 Z_{1i}$ results in the continuation ratio model assumed for the two-sample simulation study. For the three-group simulation study, groups two and three are assumed to have the same distribution (e.g., a control group and two exchangeable treatments), $\beta_1 = \beta_2 = \beta$. For each group, censoring intervals were simulated given T. The pattern-mixture restrictions in Section 3 and the distribution of T for group g are not enough to fully identify the group-specific distribution of the censoring intervals given T. The number of free parameters in this distribution for each group is $\frac{(M+2)(M+1)}{2}$, the number of intervals minus the number of event times. These parameters (interval probabilities) were fixed at values satisfying the constraints $P_g(T = t) > P_g(T = t, L = l, R = r)$, for $g = 1, \ldots, G$. The strict inequality allows positive probability for each combination of l and r including t. The remaining $M + 2$ interval probabilities were identified from the constraints $\sum_{l \leq r} P_g(L = l, R = r) =$

22

1 and $\sum_{\{l,r\}:\,l\leq t\leq r} P_g(T = t \mid L = l, R = r)P_g(L = l, R = r) = P_g(T = t)$.

True event times, T, were drawn given g, with $\beta \in \{0, 0.75\}$ and $\boldsymbol{\theta} = \{-0.65, -0.55, -0.45, -0.15, -0.05\}$. Let $\boldsymbol{\phi} = \{\phi_g : g = 1, \ldots, G\}$ be the vector of group-specific censoring bias parameters for the censoring bias function in Equation (3.7). The true censoring bias parameters were combinations of $\{-\log(2), 0, \log(2)\}$. In this study, $\phi_2 = \phi_3$.

The empirical sizes of the tests were estimated assuming $\boldsymbol{\beta} = 0$. Empirical power was estimated for the alternative hypothesis $H_1 : S_g(\cdot) \neq S_{g'}(\cdot)$ when $\boldsymbol{\beta} = 0.75$, where $g = 1$ and $g' = 2$ for the two-sample test and where $g \neq g'$ for some g, $g' \in \{1, \ldots, G\}$ for the G-sample test. We chose $n_g = 100$, 200, 500 and performed 1000 simulations for each specification. Simulations were also performed on uncensored data. Values of the true parameters were chosen to produce between 86% and 97% censoring (i.e., $P(L \neq R)$), depending on $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$. For censored data, two weight functions were used for the IWD test, $\hat{w}(j) = 1$ and $\hat{w}(j) = w^*(j) \equiv \frac{\prod_{g=1}^{G} \hat{K}_g(j)}{\sum_{g=1}^{G}(n_g/n)\hat{K}_g(j)}$, where $\hat{K}_g(j)$ is the proportion of individuals in the group g sample whose serostatus is known in year j.

Simulation test results are shown in Tables 4 and 5. The first row of each sample size-specific study shows results for uncensored data. The first column shows the true $\boldsymbol{\phi}$ that generated the censoring intervals for simulations with censoring. The second column shows the assumed $\boldsymbol{\phi}$ for the model with censored data, either CAR or the true $\boldsymbol{\phi}$. Both tables show results for six tests: the IWD test with $w = 1$ and $w = w^*$ and the logrank test, all for $G = 2$ and $G = 3$. The empirical size results in Table 4 show the tests perform well with no censoring, and the performance improves as the sample size increases. When $\boldsymbol{\phi}$ are correctly specifiec, or when the bias for both parameters is of equal magnitude in the same direction

23

(e.g., true $\boldsymbol{\phi}$ are $\{-\log(2), -\log(2)\}$, but CAR is assumed), the tests perform well. Empirical size differs most from nominal size when $\boldsymbol{\phi}$ are biased in different directions (e.g., true $\boldsymbol{\phi}$ are $\{-\log(2), \log(2)\}$, but CAR is assumed). No single test performs uniformly better than the others, however, the logrank test tends to be more anticonservative than the IWD test for smaller samples sizes, even with no censoring. When the data are censored, the three-group IWD test produces the most conservative (n=500, true $\boldsymbol{\phi} = \{-\log(2), -\log(2)\}$) and most anticonservative (n=200, true $\boldsymbol{\phi} = \{-\log(2), 0\}$) results. Empirical power is shown in Table 5. In general, the test with weight w^* is more powerful than the analogous test with $w = 1$. With no censoring, the logrank test is more powerful than the IWD test. However, with censoring in smaller sample sizes, the IWD test tends to be more powerful. In larger sample sizes, the difference is negligible. The true underlying distribution has hazard ratio 2.12 ($\exp\{0.75\}$) comparing groups 2 and 3 to group 1. When groups 2 and 3 are biased to have greater (lower) hazards relative to group 1, power is increased (decreased).

6   Appendix 2: Bayesian Algorithm

The Bayesian algorithm is a G-group version of that described in Shardell *et al.* (2006). Let I_g be complete data and $\boldsymbol{\omega}_g$ be observed data for all individuals in group g. First, starting values are chosen for censoring bias parameters, $\boldsymbol{\phi}_{(0)}$, and event-time probabilities, $\boldsymbol{p}^{(0)}$. The algorithm proceeds by simulating quantities in three steps for iteration $s = 1, \ldots, N_{sim}$:

1. Simulate $I_g^{(s)}$ from $p(I_g \mid \boldsymbol{\omega}_g, \boldsymbol{p}_g^{(s-1)}, \boldsymbol{\phi}^{(s-1)})$.

2. Simulate $\boldsymbol{p}_g^{(s)}$ from $p(\boldsymbol{p}_g \mid \boldsymbol{\omega}_g, \boldsymbol{\phi}^{(s-1)}, I_g^{(s)}) = p(\boldsymbol{p}_g \mid I_g^{(s)})$.

3. Simulate $\boldsymbol{\phi}^{(s)}$ from $p(\boldsymbol{\phi} \mid \boldsymbol{\omega}_g, I_g^{(s)}, \boldsymbol{p}_g^{(s)})$,

where $p(\cdot)$ denotes the density.

The vector of imputed event indicators, $I_{ig}^{(s)}$, for person i in group g are simulated from a truncated multinomial distribution,

$$
\prod_{j=1}^{M+1} \left[\frac{\omega_{igj} p_{gj}^{(s-1)} \exp\{q_g(j, l, r, \delta, \boldsymbol{\phi}^{(s-1)})\}}{\sum_{k=1}^{M+1} \omega_{igk} p_{gk}^{(s-1)} \exp\{q_g(k, l, r, \delta, \boldsymbol{\phi}^{(s-1)})\}} \right]^{I_{igj}}.
$$

The $I_g^{(s)}$ are aggregated into group-by-time frequencies. Let $n_{gj}^{(s)}$ denote the simulated event count during interval j among those in group g at iteration S, and $\boldsymbol{n}_g^{(s)} = \{n_{g1}^{(s)} \ldots n_{g(M+1)}^{(s)}\}$. Conditional on $\boldsymbol{n}_g^{(s)}$, $\boldsymbol{p}_g$ is independent of $\boldsymbol{\phi}$ and $\boldsymbol{\omega}_g$. Therefore, $\boldsymbol{p}_g^{(s)}$ can be simulated in Step 2 from $p(\boldsymbol{p}_g \mid I_g^{(s)}) = p(\boldsymbol{p}_g \mid \boldsymbol{n}_g^{(s)})$, a Dirichlet distribution with $\boldsymbol{\alpha}_g$ in Equation (3.8) replaced by $\boldsymbol{\alpha}_g + \boldsymbol{n}_g^{(s)}$. The $\boldsymbol{\phi}$ are simulated in Step 3 via a Metropolis-Hastings step. Let $\boldsymbol{I}^{(s)}$ denote the iteration 4 vector of simulated event indicators and $\boldsymbol{\omega}$ denote observed data across all groups. The candidate, $\boldsymbol{\phi}^*$, is simulated from the jumping distribution at iteration s, $J_s(\boldsymbol{\phi}^* \mid \boldsymbol{\phi}^{(s-1)})$, and is accepted with probability $\min(1, r_{MH})$, where

$$
\frac{p(I^{(s)} \mid \boldsymbol{\omega}, \boldsymbol{p}_g^{(s)}, \boldsymbol{\phi}^*) p(\boldsymbol{\phi}^*) J_s(\boldsymbol{\phi}^{(s-1)} \mid \boldsymbol{\phi}^*)}{p(I^{(s)} \mid \boldsymbol{\omega}, \boldsymbol{p}_g^{(s)}, \boldsymbol{\phi}^{(s-1)}) p(\boldsymbol{\phi}^{(s-1)}) J_s(\boldsymbol{\phi}^* \mid \boldsymbol{\phi}^{(s-1)})}.
$$

Acknowledgments

helpful discussions.

References

[1] AKRITAS, M.G. (1988). Rank tests with interval-censored data. *The Annals of Statistics* 16, 1490-1502.

[2] BARNDORFF-NIELSEN, O.E. AND COX, D.R. (1989). *Asymptotic Techniques for Use in Statistics.* London: Chapman and Hall.

[3] BEBCHUK, J.D. AND BETENSKY, R.A. (2000). Multiple imputation for simple estimation of the hazard function based on interval censored data. *Statistics in Medicine* 19, 405-419.

[4] BIRMINGHAM, J., ROTNITZKY, A. AND FITZMAURICE, G.M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society, Series B, Methodological* 65, 275-297.

[5] CALLE, M.L. AND GOMEZ, G. (2001). Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference* 96, 73-87.

[6] CHI, Y. (2001). Simulation study and implementation of the tests based on weighted Turnbull's estimators for interval-censored data. *Statistics in Medicine*, 20, 281-294.

[7] CHOU, R., HUFFMAN, L.H., FU, R., SMITS, A.K. AND KORTHUIS, P.T. (2005). Screening for HIV: A review of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine* 13, 55-73.

[8] COWLES, K.P. AND CARLIN, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* 91, 883-904.

[9] DEMPSTER, P., LAIRD, N. AND RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological* 39, 1-22.

[10] FANG, H.B., SUN, J. AND LEE, M.T. (2002). Nonparametric survival comparisons for interval-censored continuous data. *Statistica Sinica* 12, 1073-1083.

[11] FAY, M.P. (1996). Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics* 52, 811-822.

[12] FAY, M.P. (1999). Comparing several score tests for interval censored data (Corr: 1999V18 p2681). *Statistics in Medicine* 18, 273-285.

[13] FINKELSTEIN, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* 42, 845-854.

[14] FRIEDMAN, S.R., JOSE, B., DEREN, S., DES JARLAIS, D.C. AND NEAI-GUS, A. (1995). Risk factors for human immunodeficiency virus seroconversion among out-of-treatment drug injectors in high and low seroprevalence cities. The National AIDS Research Consortium. *American Journal of Epidemiology* 142, 864-874.

[15] GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions in Pattern Analysis and Machine Intelligence* 6 721-741.

[16] GILL, R.D., VAN DER LAAN, M.J. AND ROBINS, J.M. (1997). Coarsening at random: characterizations, conjectures and counter-examples. In Lin, D.Y. and Fleming, T.R. (eds), *State of the Art in Survival Analysis*. New York: Springer Lecture Notes in Statistics 123, pp. 255-294.

[17] HASTINGS, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97-109.

[18] HEITJAN, D.F. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics* 49, 1099-1109.

[19] HEITJAN, D.F. AND RUBIN, D.B (1991). Ignorability and coarse data. *The Annals of Statistics* 19, 2244-2253.

[20] KADANE, J.B. (1993). Subjective Bayesian analysis for surveys with missing data. *The Statistician* 42, 415-426.

[21] KAPLAN, E.L. AND MEIER, P. (1958). Non-Parametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457-481.

[22] LOUIS, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological* 44, 226-233.

[23] MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 50, 163-170.

[24] NELSON, K.E., GALAI, N., SAFAIEAN, M., STRATHDEE, S.A., CELEN-TANO, D.D. AND VLAHOV, D. (2002). Temporal trends in the incidence of human immunodeficiency virus infection and risk behavior among injection drug users in Baltimore, Maryland, 1988-1998. *American Journal of Epidemiology* 56, 641-653.

[25] PAN, W. (2000a). A two-sample test with interval censored data via multiple imputation. *Statistics in Medicine* 19, 1-11.

[26] PAN, W. (2000b). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* 56, 199-203.

[27] PETRONI, G.R. AND WOLFE, R.A. (1994). A two-sample test for stochastic ordering with interval-censored data. *Biometrics* 50, 77-87.

[28] SHARDELL, M., SCHARFSTEIN, D.O., AND BOZZETTE, S.A. (2006). Survival curve estimation for informatively coarsened discrete event-time data. *Statistics in Medicine* doi:10.1002/sim.2697.

[29] STRATHDEE, S.A., GALAI, N., SAFAIEAN, M., CELENTANO, D.D., VLAHOV, D., JOHNSON, L. AND NELSON, K.E. (2001). Sex differences in risk factors for HIV seroconversion among injection drug users: a 10-year perspective. *Archives of Internal Medicine* 161, 1281-1288.

[30] SUN, J. (1996). A non-parametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine* 15, 1387-1395.

[31] TANNER, M.A. AND WANG, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82, 528-540.

[32] TURNBULL, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B, Methodological* 38, 290-295.

[33] VLAHOV, D., ANTHONY, J.C., MUNOZ, A., MARGOLICK, J., NELSON, K.E., CELENTANO, D.D., SOLOMON, L. AND POLK, B.F. (1991). The ALIVE study, a longitudinal study of HIV-1 infection in intravenous drug users: description of methods and characteristics of participants. *NIDA Research Monographs* 109, 75-100.

[34] WIEBEL, W. AND ALTMAN, N. (1988). AIDS prevention outreach to IV-DUs in four US cities, poster presentation at the Fourth International Conference on AIDS, Stockholm.

28

[35] ZHAO, Q. AND SUN, J. (2004). A generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine* 23, 1621-1629.

Figure 1: ALIVE schematic used to elicit expert information.

a.

b.

c.

29
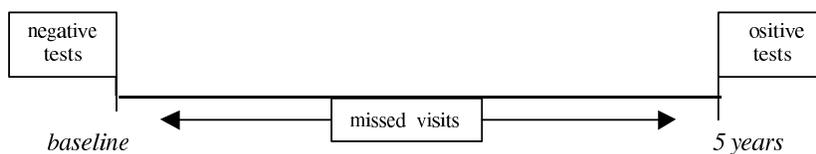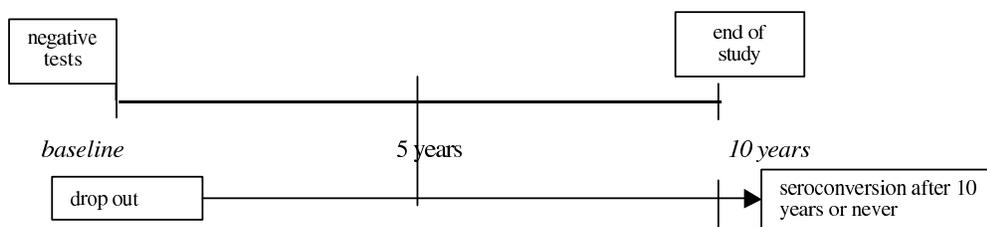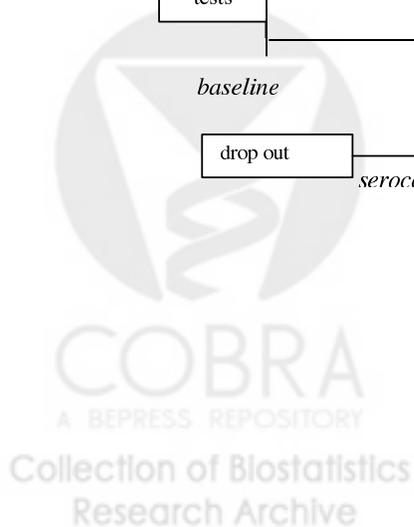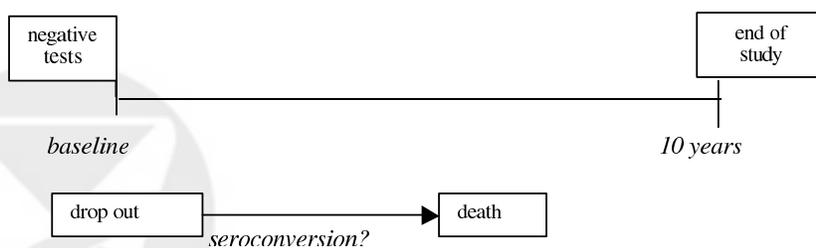
Figure 2: Correlation elicitation device for $\exp\{\phi_{g1}\}$. Axes are $P_g(L = 1, \ R = 5, \ \Delta = 0 \mid T = 5)/P_g(L = 1, \ R = 5, \Delta = 0 \mid T = 1)$ $(g = 1$, non-sharers; $g = 2$, needle-sharers). ρ denotes the correlation coefficient between $\exp\{\phi_{11}\}$ and $\exp\{\phi_{21}\}$.

Needle sharers and non-sharers are independent with respect to their relative probabilities of being interval-censored (ρ=0).	Needle sharers and non-sharers are somewhat correlated with respect to their relative probabilities of being interval-censored (ρ=0.5).	Needle sharers and non-sharers are very correlated with respect to their relative probabilities of being interval-censored (ρ=0.9).

Figure 3: Elicited prior and [**posterior**] correlation matrix for $\exp\{\boldsymbol{\phi}\}$.

Censoring/Needle-sharing Group	Interval-censored Needle-sharers ϕ_{11}	Dropped-out Needle-sharers ϕ_{12}	Dead Needle-sharers ϕ_{13}	Interval-censored Non-sharers ϕ_{21}	Dropped-out Non-sharers ϕ_{22}
Dropped-out Needle-sharers ϕ_{12}	0.00 [**-0.02**]				
Dead Needle-sharers ϕ_{13}	0.00 [**-0.03**]	0.60 [**0.58**]			
Interval-censored Non-sharers ϕ_{21}	0.75 [**0.72**]	0.00 [**-0.02**]	0.00 [**-0.02**]		
Dropped-out Non-sharers ϕ_{22}	0.00 [**-0.01**]	0.75 [**0.72**]	0.45 [**0.42**]	0.00 [**-0.01**]	
Dead Non-sharers ϕ_{23}	0.00 [**-0.03**]	0.50 [**0.48**]	0.75 [**0.73**]	0.00 [**-0.02**]	0.55 [**0.52**]

Figure 4: ALIVE Bayesian results. Posterior (solid line, needle sharers; dashed line, non-sharers) and prior (dotted line) densities of one-year, five-year, and ten-year seronegative probabilities. Posterior (gray) and prior (white) box plots of $\exp\{\phi\}$.

32
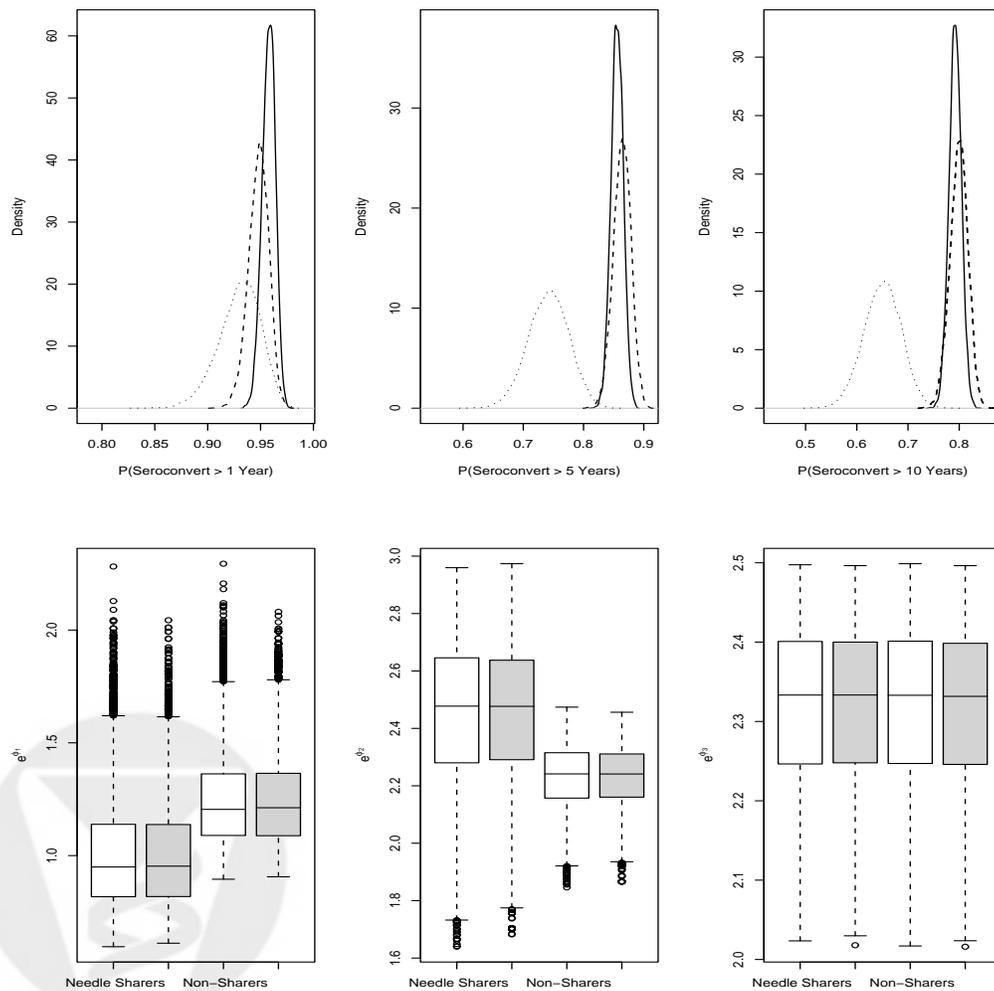
Figure 5: ALIVE Bayesian comparisons. Posterior density for $Z(\boldsymbol{p})$ (solid line) with standard normal kernel (dotted line). Mean posterior (solid line, needle sharers; dashed line, non-sharers) and prior (dotted line) survival curves with posterior tail probability.
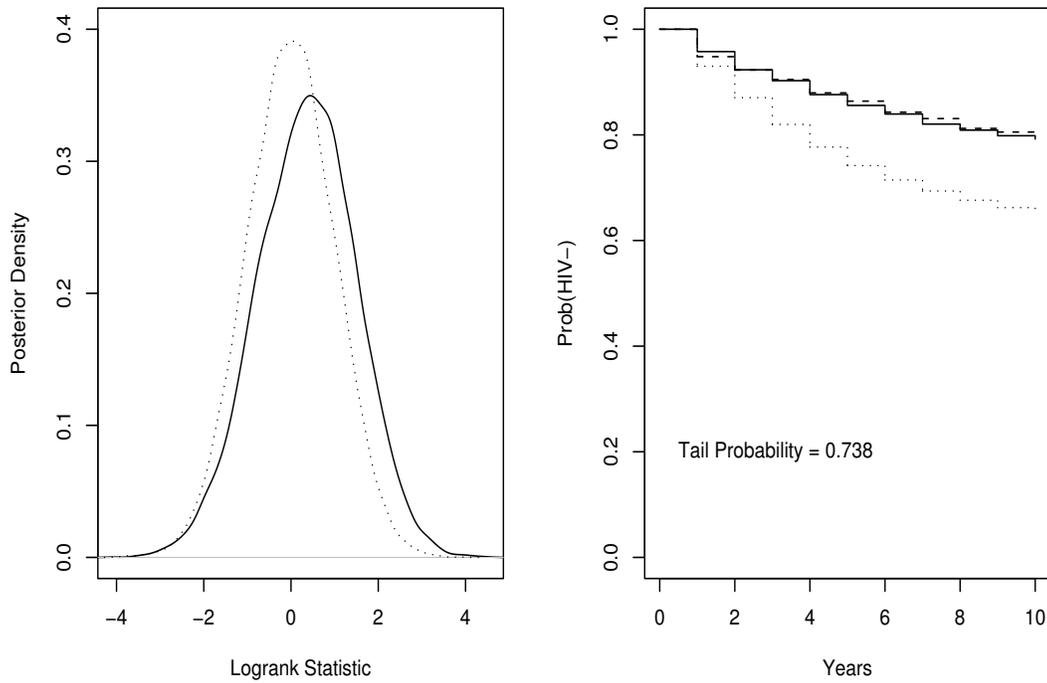
Table 1: ALIVE frequentist results. Survival estimates (95% confidence intervals) at years 1, 5, and 10.

Group	Years	CAR	Max ϕ_1, Min ϕ_2	Min ϕ_1, Max ϕ_2
Non-sharers				
	1	0.930 (0.896, 0.953)	0.953 (0.929, 0.969)	0.945 (0.918, 0.963)
	5	0.838 (0.795, 0.873)	0.887 (0.855, 0.912)	0.875 (0.840, 0.902)
	10	0.767 (0.718, 0.809)	0.829 (0.790, 0.861)	0.813 (0.772, 0.848)
Needle-sharers				
	1	0.949 (0.927, 0.964)	0.953 (0.935, 0.967)	0.968 (0.953, 0.978)
	5	0.825 (0.797, 0.850)	0.856 (0.832, 0.877)	0.879 (0.858, 0.897)
	10	0.755 (0.723, 0.784)	0.793 (0.765, 0.819)	0.822 (0.797, 0.844)

Table 2: Elicited hyperparameters for beta distributions used in Bayesian analyses.

Needle-sharing	Censoring	ϕ	Shape	Scale
Yes	Interval-censored	ϕ_{11}	2.00	7.75
	Dropped out	ϕ_{12}	5.25	2.75
	Dead	ϕ_{13}	2.00	1.00
No	Interval-censored	ϕ_{21}	2.00	9.50
	Dropped out	ϕ_{22}	3.75	2.00
	Dead	ϕ_{23}	2.00	1.00

Table 3: ALIVE Bayesian results. Mean posterior survival probabilities (95% credible intervals) at years 1, 5, and 10.

Group	Years	CAR	Max ϕ_1, Min ϕ_2	Min ϕ_1, Max ϕ_2	Random ϕ
Non-sharers					
	1	0.930 (0.903, 0.953)	0.951 (0.931, 0.967)	0.944 (0.923, 0.962)	0.948 (0.928, 0.965)
	5	0.820 (0.784, 0.854)	0.868 (0.837, 0.894)	0.855 (0.824, 0.884)	0.864 (0.834, 0.891)
	10	0.745 (0.704, 0.784)	0.803 (0.768, 0.837)	0.789 (0.751, 0.824)	0.799 (0.763, 0.832)
Needle-sharers					
	1	0.945 (0.930, 0.960)	0.951 (0.937, 0.964)	0.963 (0.952, 0.973)	0.958 (0.945, 0.969)
	5	0.811 (0.785, 0.835)	0.839 (0.817, 0.860)	0.862 (0.842, 0.880)	0.856 (0.836, 0.875)
	10	0.736 (0.708, 0.763)	0.771 (0.745, 0.795)	0.798 (0.775, 0.821)	0.792 (0.768, 0.816)

Table 4: Simulation results: empirical size of IWD and LR tests. 1000 iterations, $\beta = 0$, $M = 4$, $n_1 = n_2 = n_3 = n_.$, 86% censoring. $\alpha = 0.05$.

True ϕ	Modeled ϕ	2 groups $w = 1$	2 groups w^*	2 groups LR	3 groups $w = 1$	3 groups w^*	3 groups LR
n. = 100							
no censoring	Truth	0.054	–	0.061	0.055	–	0.064
$-\log(2), -\log(2)$	Truth	0.043	0.038	0.048	0.043	0.042	0.059
	CAR	0.055	0.056	0.066	0.063	0.060	0.075
$-\log(2), 0$	Truth	0.050	0.058	0.048	0.052	0.052	0.057
	CAR	0.129	0.127	0.131	0.128	0.131	0.137
$-\log(2), \log(2)$	Truth	0.058	0.059	0.062	0.050	0.054	0.061
	CAR	0.311	0.292	0.297	0.311	0.306	0.304
n. = 200							
no censoring	Truth	0.054	–	0.055	0.055	–	0.059
$-\log(2), -\log(2)$	Truth	0.044	0.044	0.036	0.049	0.054	0.051
	CAR	0.060	0.059	0.059	0.054	0.058	0.054
$-\log(2), 0$	Truth	0.055	0.052	0.051	0.072	0.071	0.066
	CAR	0.208	0.201	0.209	0.197	0.198	0.204
$-\log(2), \log(2)$	Truth	0.053	0.052	0.058	0.052	0.055	0.050
	CAR	0.500	0.490	0.464	0.528	0.518	0.511
n. = 500							
no censoring	Truth	0.050	–	0.049	0.052	–	0.054
$-\log(2), -\log(2)$	Truth	0.054	0.052	0.053	0.062	0.062	0.061
	CAR	0.045	0.043	0.050	0.061	0.059	0.059
$-\log(2), 0$	Truth	0.054	0.056	0.056	0.059	0.060	0.063
	CAR	0.389	0.377	0.366	0.364	0.364	0.365
$-\log(2), \log(2)$	Truth	0.047	0.048	0.049	0.034	0.035	0.040
	CAR	0.891	0.880	0.863	0.917	0.909	0.894

Table 5: Simulation results: empirical power of IWD and LR tests. 1000 iterations, $\beta = 0.75$, $M = 4$, $n_1 = n_2 = n_3 = n_.$, 97% censoring when $Z = 1$ and 86% censoring when $Z = 0$. $\alpha = 0.05$.

True ϕ	Modeled ϕ	2 groups $w=1$	2 groups $w*$	2 groups LR	3 groups $w=1$	3 groups $w*$	3 groups LR
$n_. = 100$							
no censoring		0.966	–	0.972	0.976	–	0.982
$-\log(2),\,-\log(2)$	Truth	0.860	0.860	0.850	0.909	0.900	0.907
	CAR	0.638	0.668	0.601	0.697	0.723	0.684
$-\log(2),\,0$	Truth	0.810	0.837	0.785	0.867	0.884	0.865
	CAR	0.929	0.939	0.912	0.970	0.976	0.963
$-\log(2),\,\log(2)$	Truth	0.756	0.784	0.750	0.820	0.841	0.823
	CAR	1.000	1.000	0.999	1.000	1.000	1.000
$n_. = 200$							
no censoring		0.999	–	0.999	1.000	–	1.000
$-\log(2),\,-\log(2)$	Truth	0.993	0.989	0.991	0.998	0.998	0.999
	CAR	0.916	0.931	0.893	0.953	0.968	0.941
$-\log(2),\,0$	Truth	0.973	0.979	0.971	0.992	0.994	0.988
	CAR	0.996	0.997	0.996	0.999	0.999	0.997
$-\log(2),\,\log(2)$	Truth	0.961	0.973	0.959	0.982	0.991	0.985
	CAR	1.000	1.000	1.000	1.000	1.000	1.000
$n_. = 500$							
no censoring		1.000	–	1.000	1.000	–	1.000
$-\log(2),\,-\log(2)$	Truth	1.000	1.000	1.000	1.000	1.000	1.000
	CAR	1.000	1.000	1.000	1.000	1.000	1.000
$-\log(2),\,0$	Truth	1.000	1.000	1.000	1.000	1.000	1.000
	CAR	1.000	1.000	1.000	1.000	1.000	1.000
$-\log(2),\,\log(2)$	Truth	1.000	1.000	1.000	1.000	1.000	1.000
	CAR	1.000	1.000	1.000	1.000	1.000	1.000