UC Irvine
UC Irvine Previously Published Works

Title

Deductive Semiparametric Estimation in Double-Sampling Designs with Application to PEPFAR

Permalink

https://escholarship.org/uc/item/7wv9f3qm

Journal

Statistics in Biosciences, 12(3)

ISSN

1867-1764

Authors

Qian, Tianchen
Frangakis, Constantine
Yiannoutsos, Constantin

Publication Date

2020-12-01

DOI

10.1007/s12561-019-09262-2

Copyright Information

Peer reviewed

Deductive semiparametric estimation in Double-Sampling Designs with application to PEPFAR

Tianchen Qian[*]   Constantine Frangakis[†]   Constantin Yiannoutsos[‡]

June 27, 2019

Abstract

Non-ignorable dropout is common in studies with long follow-up time, and it can bias study results unless handled carefully in the study design and the statistical analysis. A double-sampling design allocates additional resources to pursue a subsample of the dropouts and find out their outcomes, which can address potential biases due to non-ignorable dropout. It is desirable to construct semiparametric estimators for the double-sampling design because of their robustness properties. However, obtaining such semiparametric estimators remains a challenge due to the requirement of the analytic form of the efficient influence function (EIF), the derivation of which can be *ad hoc* and difficult for the double-sampling design. Recent work has shown how the derivation of EIF can be made deductive and computerizable using the functional derivative representation of the EIF in nonparametric models. This approach, however, requires deriving the mixture of a continuous distribution and a point mass, which can itself be challenging for complicated problems such as the double-sampling design. We propose semiparametric estimators for the survival probability in double-sampling designs by generalizing the deductive and computerizable estimation approach. In particular, we propose to build the semiparametric estimators based on a *discretized support* structure, which approximates the possibly continuous observed data distribution and circumvents the derivation of the mixture distribution. Our approach is *deductive* in the sense that it is expected to produce semiparametric locally efficient estimators within finite steps without knowledge of the EIF. We apply the proposed estimators to estimating the mortality rate in a double-sampling design component of the President's Emergency Plan for AIDS Relief (PEPFAR) program. We evaluate the impact of double-sampling selection criteria on the mortality rate estimates. Simulation studies are conducted to evaluate the robustness of the proposed estimators.

---

[*]Department of Statistics, Harvard University. qiantianchen@fas.harvard.edu
[†]Department of Biostatistics, Johns Hopkins University.
[‡]Department of Biostatistics, Indiana University.

1 Introduction

Studies with long follow-up often suffer from a high dropout rate. Dropouts can depend on the outcome of interest, even after adjusting for observed covariates. This makes the dropouts "non-ignorable" and biases the analysis based solely on the non-dropouts (Rubin, 1976). As a way to handle non-ignorable dropouts, double-sampling designs allocate additional resources to pursue a sample of the dropouts and find out their outcomes (Baker et al., 1993; Glynn et al., 1993; Frangakis and Rubin, 2001; Cochran, 2007). Compared to standard double-sampling (where each dropout is sampled with equal probability), the double-sampling can be more practical or informative if it targets dropouts whose history at the dropout time has specific profiles. For example, An et al. (2009) found that such "profile designs" can save more than 35% of resources compared to the standard double-sampling design. It can also be more practical to double-sample relatively recent dropouts. For data analysis of such profile double-sampling designs, An et al. (2014) employed a parametric approach to estimate the survival probability. However, analyses of such designs can be more reliable if they do not rely heavily on parametric assumptions. Semiparametric estimators (Newey, 1990; Tsiatis, 2007), which focus on modeling the parameter of interest and treat the rest of the model as nuisance parameters, have had great success in various areas (Cox, 1972; Liang and Zeger, 1986; Robins et al., 1994; Zhang et al., 2008), and they are a promising alternative to the parametric estimators for double-sampling designs. Robins et al. (2001) had suggested a possible way of deriving a semiparametric estimator for double-sampling designs, but that and any other such existing proposals rely on first coming up with and then verifying conjectures "by hand". Such a process is prone to human errors and is not generalizable.

Recently, Frangakis et al. (2015) proposed a different approach to construct semiparametric estimators. Contrary to the classical semiparametric framework (Bickel et al., 1993; Tsiatis, 2007) which relies heavily on the conjecture and verification of the analytic form of the efficient influence function (EIF), their approach produces semiparametric locally efficient estimators by utilizing the Gateaux derivative representation of the EIF in nonparametric models. This estimation procedure is *deductive*, in the sense that the estimator is computed through a discrete and finite set of steps, without requiring conjecturing for or theoretically deriving the EIF. If a semiparametric estimator can be obtained deductively, human effort can be saved from difficult mathematical derivations and can be transferred, for example, to designing new studies. Their approach is not directly applicable to the data

analysis of double-sampling designs, however, because their estimation procedure requires analytically evaluating the parameter of interest at a mixture of a continuous distribution and a point mass. This analytic step is feasible for certain problems (e.g., for the two-phase design considered in Frangakis et al. (2015)), but can become highly error-prone when derived by hand in complicated problems like estimation of the survival probability in the double-sampling design.

In this paper, we develop semiparametric estimators for the survival probability in double-sampling designs, by generalizing the deductive estimator by Frangakis et al. (2015) to problems where the aforementioned analytic evaluation is infeasible. We use a *discretized support* structure to circumvent the analytical evaluation of the parameter of interest at a mixture distribution. The discretized support is a discretized approximation of the sample space constructed from the observed data, which enables numerical calculation of the estimand at the mixture. A similar discretization idea was used in Chamberlain (1987) for deriving efficiency bounds in nonparametric models. Our method produces semiparametric locally efficient estimators within finite steps without knowledge of the EIF; i.e., it is deductive.

We apply the method to estimating the mortality rate using data from a double-sampling design component of the President's Emergency Plan for AIDS Relief (PEPFAR), an HIV monitoring and treatment program (Geng et al., 2015). In addition, we explore and discuss how certain restrictions on the double-sampling selection criteria may impact the estimated mortality rate, as the double-sampling for this study is more pragmatic if restricted to relatively recent dropouts (An et al., 2014).

The paper is organized as follows. In Section 2, we introduce the double-sampling design and the parameter of interest. In Section 3, we present the proposed estimation method for survival probability in double-sampling designs. In particular, we present and discuss how the discretized support idea facilitates the deductive estimation in double-sampling designs. In Section 4, we present results from simulation studies which show the performance of the proposed estimators and several other existing methods. In Section 5, we apply the proposed method to estimating the mortality rate using data from a double-sampling design component of PEPFAR, and we discuss the impact of restrictions on the double-sampling selection criteria on the estimates. Section 6 concludes with discussion.

2 Double-sampling design and parameter of interest

2.1 Double-sampling design

For clarity, we present arguments with a double-sampling design as shown in Figure 1, which illustrates different patient types in the double-sampling component of PEPFAR. This figure is a modified version of Figure 1 in An et al. (2014). First, we describe characteristics that are inherent to a patient (i.e., potential outcomes (Rubin, 1976)), which would be realized if the program could enroll and follow a patient indefinitely with a standard effort. Then, we describe the actual double-sampling design consisting of two phases and the observed data.

Patient inherent characteristics and parameter of interest. For each patient, there is an enrollment time E_i, covariates Z_i at enrollment, and a survival time T_i (time from enrollment to death). The parameter of interest is $P(T_i > t)$ for a given t, which is the proportion of patients surviving beyond year t for a population represented as a random sample by the enrolled patients. If the program were to follow patients indefinitely with a standard effort while they are alive, some patients would discontinue contact from the monitoring (patient types (b1), (c), (d), (e1), and (e2) in Figure 1). For these patients, labeled true dropouts and indicated by $R_i = 0$, denote by L_i the time from enrollment to dropout and denote by W_i any information available at dropping out in addition to Z_i. For instance, W_i may include L_i and some longitudinal measurements.

Phase 1 of the double-sampling design and missingness due to administrative censoring. In the first phase, the actual design enrolls patients at times E_i and monitors them with standard effort, not indefinitely, but until "Now"—the present time. The time from enrollment E_i to "Now" is the time to administrative censoring, denoted by C_i. For patient i, define $\Delta_i = 1$ if $C_i \geq T_i$ and $\Delta_i = 0$ otherwise. Define $X_i = \min(C_i, T_i)$.

Simply based on Phase 1, the standard survival data (X_i, Δ_i) are not observed for dropouts whose time to dropout satisfies $L_i < X_i$. Denote such *observed* dropout patients by $R_i^{\mathrm{obs}} = 0$ (patient types (c), (d), (e1), and (e2) in Figure 1), and denote the rest (*observed* non-dropouts) by $R_i^{\mathrm{obs}} = 1$.

Phase 2: information after double-sampling. Phase 2 of the design selects a subset of the observed dropouts, called a double-sample, and allocates additional resources for searching for and finding this double-sample at "Now". Denote by S_i the indicator of being selected

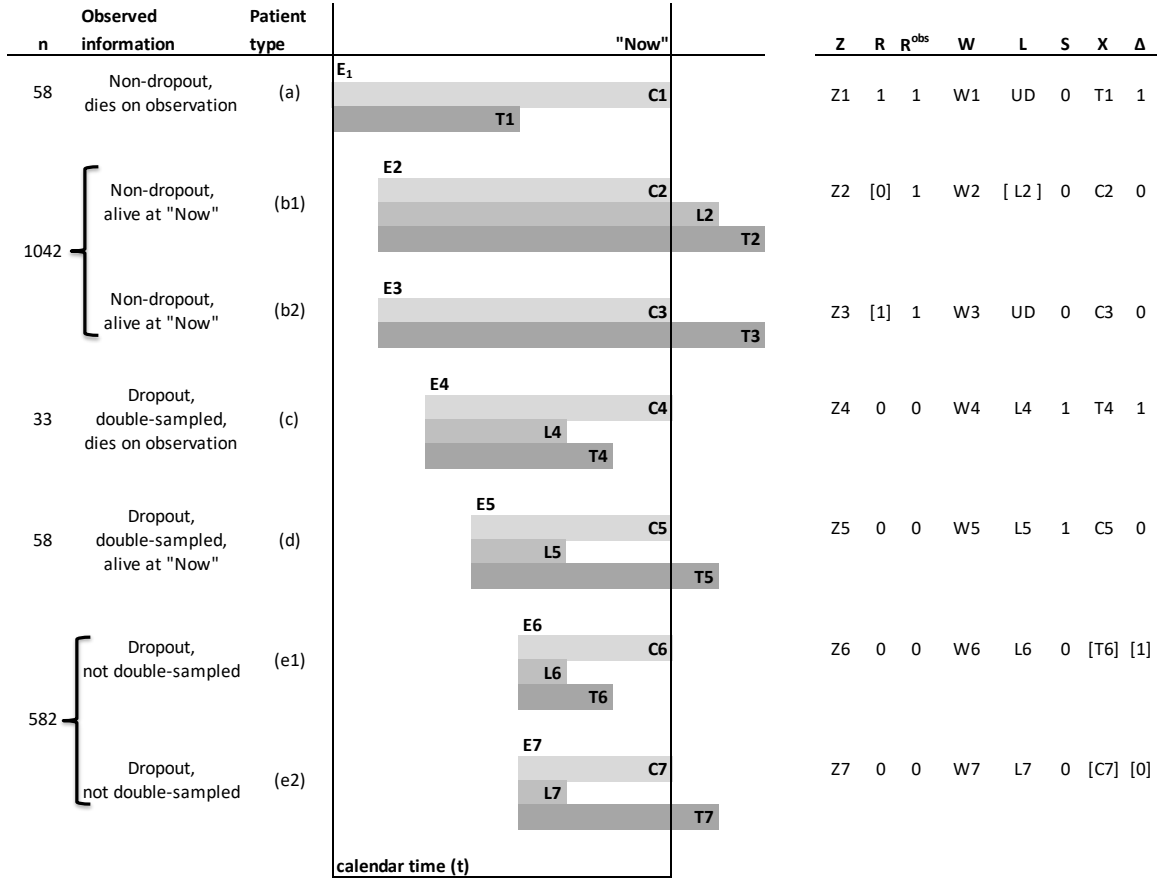| n	Observed information	Patient type	Z	R	R^{obs}	W	L	S	X	Δ
58	Non-dropout, dies on observation	(a)	Z1	1	1	W1	UD	0	T1	1
1042	Non-dropout, alive at "Now"	(b1)	Z2	[0]	1	W2	[L2]	0	C2	0
	Non-dropout, alive at "Now"	(b2)	Z3	[1]	1	W3	UD	0	C3	0
33	Dropout, double-sampled, dies on observation	(c)	Z4	0	0	W4	L4	1	T4	1
58	Dropout, double-sampled, alive at "Now"	(d)	Z5	0	0	W5	L5	1	C5	0
582	Dropout, not double-sampled	(e1)	Z6	0	0	W6	L6	0	[T6]	[1]
	Dropout, not double-sampled	(e2)	Z7	0	0	W7	L7	0	[C7]	[0]

Figure 1: Characteristics for different patient types in a double-sampling design. E_i denotes the enrollment time; C_i denotes the time to administrative censoring; T_i denotes the time to death; L_i denotes the time to dropout. Z denotes covariates observed at enrollment; R denotes the true dropout status; R^{obs} denotes the observed dropout status; W denotes the dropout time and some other longitudinal measure; S denotes the double-sampling selection indicator; (X, Δ) is the standard survival outcome: $X = \min(T, C)$, $\Delta = \mathbb{1}(T \leq C)$, where $\mathbb{1}(\cdot)$ is the indicator function. "UD" means undefined. Quantities in brackets are unobserved. Note that unobserved L and undefined L cannot be distinguished from the observed data. Column n denotes the number of patients of each type in the double-sampling data set of PEPFAR. The figure is a modified version of Figure 1 in An et al. (2014).

as a double-sample. The survival data (X_i, Δ_i) is observed for the double-samples (those with $S_i = 1$).

The observed data for patient i is $(C_i, R_i^{\text{obs}}, Z_i, W_i, S_i, X_i, \Delta_i)$. For the patients with $R_i^{\text{obs}} = 1$, we define $S_i = 0$. For the patients with $R_i^{\text{obs}} = 0$ and $S_i = 0$, the survival data (X_i, Δ_i) is not observed, and we define $(X_i, \Delta_i) = (\text{NA}, \text{NA})$. If W_i only includes the dropout time L_i, W_i is undefined for the patients with $R_i^{\text{obs}} = 1$ and we denote it by NA; if W_i includes additional longitudinal measurement that is observed for those $R_i^{\text{obs}} = 1$, then W_i is defined for all patients. We use $\perp$ to denote independence between random variables.

2.2   Identification of parameter

To identify the parameter of interest $P(T_i > t)$ by the observed data distributions, we make the following assumptions.

Assumption 1 (Patient-dependent double-sampling). *For each observed dropout, S_i can depend on the patient characteristics (Z_i, W_i); after we condition on (Z_i, W_i), double-sampling S_i is independent of survival and enrollment times:*

$$S_i \perp (T_i, E_i) \mid R_i^{\text{obs}} = 0, Z_i, W_i.$$

Assumption 2 (Exchangeability of enrollment times). *Patients enrolled at different times (or equivalently, patients having different time to administrative censoring $C_i = \text{"Now"} - E_i$) have similar survival times after conditioning on Z_i:*

$$T_i \perp C_i \mid Z_i.$$

Assumption 1 means that the probability of double-sampling an observed dropout can depend on his/her characteristics known up to the dropout time L_i, but not depending on other unobserved information (such as prognostic information for the survival that are not included in (Z_i, W_i)). This can be ensured by including in (Z_i, W_i) the variables used by the investigators to select the double-samples. Assumption 2 means that there are no time-trend in the survival times after conditional on Z_i; this is plausible in practice when the design is limited to a window period of enrollment. Intuitively, Assumption 1 helps to recover the information missing due to dropout, and Assumption 2 helps to recover the information missing due to administrative censoring.

Throughout the paper, we use $P(\cdot)$ to denote the probability of an event or a point in the sample space, and we use $\mathrm{pr}(\cdot)$ to denote the probability density function of a random variable with with respect to an appropriate dominating measure (the Lebesgue measure for continuous variables and the counting measure for discrete variables). Under Assumptions 1 and 2, the estimand $P(T_i > t)$ is identifiable from the following components of the observed data distribution:

$$P(R_i^{\mathrm{obs}} = 1); \qquad \mathrm{pr}(Z_i \mid R_i^{\mathrm{obs}} = 1); \qquad \mathrm{pr}\{(X_i, \Delta_i) \mid Z_i, R_i^{\mathrm{obs}} = 1\}; \qquad (1)$$

$$P(R_i^{\mathrm{obs}} = 0); \qquad \mathrm{pr}(Z_i, W_i \mid R_i^{\mathrm{obs}} = 0); \qquad \mathrm{pr}\{(X_i, \Delta_i) \mid Z_i, W_i, R_i^{\mathrm{obs}} = 0\}, \qquad (2)$$

In particular, by Assumption 1 the distribution from the double-sampled individuals $\mathrm{pr}\{(X_i, \Delta_i) \mid Z_i, W_i, R_i^{\mathrm{obs}} = 0, S_i = 1\}$ is the same as the one for all dropouts, $\mathrm{pr}\{(X_i, \Delta_i) \mid Z_i, W_i, R_i^{\mathrm{obs}} = 0\}$, and so, together with the second component of (2) gives, upon averaging over W_i, $\mathrm{pr}\{(X_i, \Delta_i), Z_i \mid R_i^{\mathrm{obs}} = 0\}$. That, together with (1), gives $\mathrm{pr}\{(X_i, \Delta_i), Z_i\}$. Denote by $\mathrm{surv}_t(\cdot)$ the function that takes $\cdot$ as an arbitrary distribution of (X_i, Δ_i) from independent survival and censoring times and returns the survival probability beyond t (this function is the common probability limit of the Nelson-Aalen and Kaplan-Meier estimators). Then, by Assumption 2, the estimand can be expressed in terms of the observed data distribution as

$$\tau = P(T_i > t) = \int \mathrm{surv}_t[\mathrm{pr}\{(X_i, \Delta_i) \mid Z_i\}]\mathrm{pr}(Z_i)d(Z_i). \qquad (3)$$

A detailed derivation of (3) is given in Appendix B.

3 Method

3.1 Review of the estimation procedure by Frangakis et al. (2015)

An estimand can be viewed as a functional of a distribution on the data. For example, in the double-sampling design described above, (3) shows that $P(T_i > t)$ is a functional $\tau(F)$ of the distribution F of the observed data. Estimation of $\tau(F)$ requires modeling assumptions because (3) involves regressions that, in practice, cannot be estimated fully nonparametrically. In particular, suppose the estimand τ in (3) has a nonparametric efficient influence function (EIF) denoted by $\phi(O_i, [F - \tau, \tau])$, where O_i represents the observed data from subject i and $F - \tau$ represents the remaining components of the distribution other than τ.

A semiparametric estimator $\hat{\tau}$ can be obtained by solving

$$\sum_i \phi\{O_i, [(F-\tau), \tau]\} = 0 \tag{4}$$

after substituting for $(F-\tau)$ estimates of a working model. Such an estimator is consistent and locally efficient if the working estimators of $(F-\tau)$ are consistent with convergence rates larger than $n^{1/4}$ (van der Vaart, 2000). In classical semiparametric theory (Bickel et al., 1993; Tsiatis, 2007), obtaining such a semiparametric estimator requires knowing the analytic form of the EIF ϕ, the derivation of which usually relies on first conjecturing for and then verifying the form of certain Hilbert spaces. Although this process has been successful in some settings (Robins et al., 1994; Zhang et al., 2008), it is not guaranteed to succeed for estimation in double-sampling designs.

An alternative approach to construct semiparametric estimators was proposed by Frangakis et al. (2015), which does not require the analytic form of ϕ. The key idea in Frangakis et al. (2015) is that, for any working distribution $F(\alpha)$ parametrized by α, one can approximate numerically the EIF ϕ as the numerical Gateaux derivative after perturbing the working distribution $F(\alpha)$ by a small mass at each observed data point. Formally, let $\mathcal{L}_{O_i}$ be a point mass at O_i, and let $F_{(O_i,\epsilon)}(\alpha) = (1-\epsilon)F(\alpha) + \epsilon\mathcal{L}_{O_i}$ be the perturbed distribution, i.e., a mixture of $F(\alpha)$ and $\mathcal{L}_{O_i}$, where $\epsilon > 0$ is a small positive number. One can approximate $\phi\{O_i, F(\alpha)\}$ numerically by $\mathsf{Gateaux}\{O_i, F(\alpha), \epsilon\}$, where

$$\mathsf{Gateaux}\{O_i, F(\alpha), \epsilon\} := \left[\tau\{F_{(O_i,\epsilon)}(\alpha)\} - \tau\{F(\alpha)\}\right]/\epsilon.$$

Then, assuming α is 1-dimensional, one can find the best working distribution parameter $\hat{\alpha}$ as one that makes zero the sum of the numerical EIFs, $\sum_i \mathsf{Gateaux}\{O_i, F(\alpha), \epsilon\}$. The estimator solving (4) approximately is then $\tau\{F(\hat{\alpha})\}$. The standard error of the estimator can be estimated by $n^{-1}\left[\sum_i \mathsf{Gateaux}\{O_i, F(\hat{\alpha}), \epsilon\}^2\right]^{1/2}$ with n denoting the sample size.

The estimator $\tau\{F(\hat{\alpha})\}$ has consistency properties beyond those of the maximum likelihood estimator for the same working model $F(\alpha)$, since the former depends only on the parts of F that are present in the nonparametric influence function EIF. For example, Frangakis et al. (2015) show that in the two-phase design, $\tau\{F(\hat{\alpha})\}$ shares the double robustness of the estimators require the form of the EIF. A general condition regarding the true distribution F_0 and the working model F for the consistency of $\tau\{F(\hat{\alpha})\}$ is stated in Appendix A.

9

3.2 Proposed method: deductive estimation with discretized support

A semiparametric estimator is *deductive*, if it does not rely on conjectures for the functional form of the EIF ϕ and it is produced in the sense of Turing (1937) (i.e., using a discrete and finite set of instructions, and, for every input, finishing in discrete finite steps). $\tau\{F(\hat{\alpha})\}$ in Section 3.1 is such an example. Deductive estimators are desired for analysis of the double-sampling designs, because they often have additional robustness properties that the maximum likelihood estimator does not have, and they do not require the difficult derivation of the EIF. However, directly application of the deductive estimation procedure in Frangakis et al. (2015) to the double-sampling design is infeasible, because of the difficulty in evaluating $\tau\{F_{(O_i,\epsilon)}(\alpha)\}$.

We generalize the deductive estimation procedure to double-sampling designs by incorporating the so-called *discretized support*. The purpose of the discretized support is to approximate possibly continuous working models (i.e., $F(\alpha)$) via discretization. Intuitively, if the support of a discretized working model contains all the observed data points, perturbing the working model by a point mass at an observed data point (i.e., $F_{(O_i,\epsilon)}(\alpha)$) would be as simple as changing the values in a probability table, and this would circumvent the analytic evaluation of $\tau\{F_{(O_i,\epsilon)}(\alpha)\}$. A similar discretization idea was used in Chamberlain (1987) for deriving nonparametric efficiency bounds.

In the following we use lower case letter to denote the realized value of random variables. Recall that the observed data from patient i in the double-sampling design is $(c_i, r_i^{\mathrm{obs}}, z_i, w_i, s_i, x_i, \delta_i)$, and we use O_i to denote the observed data from patient i excluding c_i; i.e., $O_i = (r_i^{\mathrm{obs}}, z_i, w_i, s_i, x_i, \delta_i)$. Suppose that among the n total patients there are m observed dropouts ($r_i^{\mathrm{obs}} = 0$) and m_1 patients get double-sampled ($r_i^{\mathrm{obs}} = 0, s_i = 1$), and that the data is ordered such that $r_1^{\mathrm{obs}} = r_2^{\mathrm{obs}} = \cdots = r_m^{\mathrm{obs}} = 0$, $r_{m+1}^{\mathrm{obs}} = \cdots = r_n^{\mathrm{obs}} = 1$, $s_1 = s_2 = \cdots = s_{m_1} = 1$, and $s_{m_1+1} = \cdots = s_n = 0$. The discretized support for those with $r_i^{\mathrm{obs}} = 0$ is defined as

$$\Omega_0 = \{0\} \otimes \{(z_1, w_1), \ldots, (z_m, w_m)\} \otimes \{(s_1, x_1, \delta_1), \ldots, (s_m, x_m, \delta_m)\},$$

where $\otimes$ denotes Cartesian product[1]. Each element in Ω_0 is a vector of the same length as $(R^{\mathrm{obs}}, Z, W, S, X, \Delta)$, and we have $O_i \in \Omega_0$ for $1 \leq i \leq m$. The discretized support for those

---

[1] The set operation $\{\cdot\}$ removes duplicates. In particular, $\{(z_1, w_1), \ldots, (z_m, w_m)\}$ is the set of all unique pairs (z_i, w_i) for $1 \leq i \leq m$, and $\{(s_1, x_1, \delta_1), \ldots, (s_m, x_m, \delta_m)\}$ is the set of all unique triples (s_i, x_i, δ_i) for $1 \leq i \leq m$.

with $r_i^{\mathrm{obs}} = 1$ is defined as

$$\Omega_1 = \{1\} \otimes \{(z_{m+1}, w_{m+1}), \ldots, (z_n, w_n)\} \otimes \{(s_{m+1}, x_{m+1}, \delta_{m+1}), \ldots, (s_n, x_n, \delta_n)\},$$

and similarly we have $O_i \in \Omega_1$ for $m+1 \leq i \leq n$. The overall discretized support is defined as

$$\Omega = \Omega_0 \cup \Omega_1,$$

and we have $O_i \in \Omega$ for $1 \leq i \leq n$.

The deductive estimation procedure in Frangakis et al. (2015) is now feasible for the double-sampling design by conducting all the computations on the discretized support Ω. Suppose we have a class of fitted working distributions $\hat{F}(\alpha)$ on Ω, parametrized by a 1-dimensional α. In other words, for each value of α, $\hat{F}(\alpha)$ is a discrete probability distribution on the sample space Ω. The particular form of $\hat{F}(\alpha)$ will be described in detail in Section 3.3. Because $O_i \in \Omega$ for all i, the mixture distribution $\hat{F}_{(O_i, \epsilon)}(\alpha) = (1 - \epsilon)\hat{F}(\alpha) + \epsilon \mathcal{L}_{O_i}$ with a small $\epsilon > 0$ is also a discrete probability distribution on Ω, and can be computed directly from $\hat{F}(\alpha)$ and O_i.

For any probability distribution G on Ω, it is straightforward to compute the induced distribution components (1) and (2) through averaging and normalization; details are provided in Appendix C for completeness. Therefore, the estimand $\tau(G)$ can be computed by (3) nonparametrically because the sample space of G is discrete. We compute $\tau\{\hat{F}(\alpha)\}$ and $\tau\{\hat{F}_{(O_i, \epsilon)}(\alpha)\}$ for all $1 \leq i \leq n$, and find $\hat{\alpha}$ that solves

$$\sum_{i=1}^{n} \mathsf{Gateaux}\{O_i, \hat{F}(\alpha), \epsilon\} = \sum_{i=1}^{n} \left[\tau\{\hat{F}_{(O_i, \epsilon)}(\alpha)\} - \tau\{\hat{F}(\alpha)\} \right] / \epsilon = 0 \tag{5}$$

with a small $\epsilon > 0$. The deductive estimator for $\tau = P(T > t)$ is $\hat{\tau} = \tau\{\hat{F}(\hat{\alpha})\}$. The standard error of $\hat{\tau}$ can be estimated by $n^{-1} \left[\sum_i \mathsf{Gateaux}\{O_i, \hat{F}(\hat{\alpha}), \epsilon\}^2 \right]^{1/2}$.

Although formally the estimation procedure looks similar to that in Frangakis et al. (2015), the underlying implementation is completely different. In the two-phase design considered in Frangakis et al. (2015), they were able to derive analytically the estimand evaluated at a mixture distribution, $\tau\{\hat{F}_{(O_i, \epsilon)}(\alpha)\}$, for any estimand τ. For the double-sampling design we consider here, however, deriving $\tau\{\hat{F}_{(O_i, \epsilon)}(\alpha)\}$ is complicated and error-prone because of the presence of survival outcomes. Our proposed discretized support

11

separates the computation of the mixture distribution $\hat{F}_{(O_i, \epsilon)}(\alpha)$ and the computation of the estimand for an arbitrary distribution $\tau(G)$, which makes the deductive estimation feasible for double-sampling designs.

The estimator $\hat{\tau}$ has the following robustness property. Suppose we decompose $F(\alpha) := (R, H(\alpha))$, where H isolates the components of the distribution that are modeled through α in step 3, and R is the remaining part of F. Generally, the expected EIF, $\bar{\phi}_{(R_0, H_0)}(R, H) := E_{(R_0, H_0)}\{\phi(O_i, R, H)\}$, is zero at the truth (R_0, H_0) but possibly also at other values of (R, H) (e.g., double-robustness). Under regularity conditions usually needed for consistency with estimating equations, the above estimator when taking a working model R_w should be consistent if the model $\{H(\alpha)\}$ includes a distribution H^* that satisfies condition (A.2) of Appendix A

$$\bar{\phi}_{(R_0, H_0)}(R_w, H^*) = 0 \text{ and } \tau(R_w, H^*) = \tau(R_0, H_0),$$

that is, the distribution H^* zeros out the limit EIF and gives the correct value of the estimand.

3.3   Working distributions $\hat{F}(\alpha)$ on the discretized support Ω

Here we give the form of the working distributions $\hat{F}(\alpha)$ used in Section 3.2. We first describe how a working distribution $\hat{F}$ on Ω is estimated based on the observed data, then describe how to extend it to $\hat{F}(\alpha)$, a class of working distributions parameterized by a 1-dimensional α.

In the following, let $o = (r^{\text{obs}}, z, w, s, x, \delta)$ denote an arbitrary point in Ω, let $P_{\hat{F}}(o)$ denote the probability of o under the distribution $\hat{F}$, and let $\hat{P}$ denote a discrete probability measure that is estimated from the observed distribution, which we will describe after (6). The working distribution $\hat{F}$ is given by

$$\begin{aligned} P_{\hat{F}}(r^{\text{obs}}, z, w, s, x, \delta) = &\hat{P}(R^{\text{obs}} = r^{\text{obs}})\hat{P}(Z = z, W = w \mid R^{\text{obs}} = r^{\text{obs}}) \\ &\times \hat{P}(S = s \mid R^{\text{obs}} = r^{\text{obs}}, Z = z, W = w) \\ &\times \hat{P}(X = x, \Delta = \delta \mid R^{\text{obs}} = r^{\text{obs}}, Z = z, W = w, S = s). \end{aligned} \tag{6}$$

For notation simplicity, we will omit the lowercase letters inside $\hat{P}(\cdot)$ for the rest of the section if no confusion is caused. For the terms on the right hand side of (6), $\hat{P}(R^{\text{obs}})$ and $\hat{P}(Z, W \mid R^{\text{obs}})$ are estimated by their empirical distributions. The selection probability for double-sampling, $\hat{P}(S \mid R^{\text{obs}} = 0, Z, W)$, is estimated by logistic regression. The probability

on the survival information, $\hat{P}(X,\Delta \mid R^{\mathrm{obs}},Z,W,S)$, is estimated as follows.

By construction of Ω, we have $(x,\delta)=(\mathsf{NA},\mathsf{NA})$ if and only if $r^{\mathrm{obs}}=0,s=0$, so $\hat{P}(X=\mathsf{NA},\Delta=\mathsf{NA}\mid R^{\mathrm{obs}}=0,Z,W,S=0)=1$. For $\hat{P}(X,\Delta\mid R^{\mathrm{obs}},Z,W,S)$ with $(R^{\mathrm{obs}}=0,S=1)$ or $R^{\mathrm{obs}}=1$, we consider two different approaches: (i) an estimate based on Cox proportional hazards models, and (ii) an estimate based on log-normal regressions. For (i), Cox proportional hazards models (Cox, 1972) are fitted under the working independence assumption $T\perp C\mid R^{\mathrm{obs}},Z,W,S$ to obtain

$$\hat{P}(T=x\mid R^{\mathrm{obs}}=0,Z,W,S=1) \text{ and } \hat{P}(C=x\mid R^{\mathrm{obs}}=0,Z,W,S=1) \text{ for } x\in\{x_1,\ldots,x_{m_1}\},$$

$$\hat{P}(T=x\mid R^{\mathrm{obs}}=1,Z,W,S=0) \text{ and } \hat{P}(C=x\mid R^{\mathrm{obs}}=1,Z,W,S=0) \text{ for } x\in\{x_{m+1},\ldots,x_n\}.$$

(7)

For (ii), log-normal regressions are fitted under the working independence assumption $T\perp C\mid R^{\mathrm{obs}},Z,W,S$ to obtain probability density functions for $T\mid R^{\mathrm{obs}}=0,Z,W,S=1$, $C\mid R^{\mathrm{obs}}=0,Z,W,S=1$, $T\mid R^{\mathrm{obs}}=1,Z,W,S=0$, and $C\mid R^{\mathrm{obs}}=1,Z,W,S=0$. The probability density functions are then converted to (7). For either approach, the fitted probability on (X,Δ) is then computed as

$$\hat{P}(X=x,\Delta=\delta\mid R^{\mathrm{obs}},Z,W,S)=\left\{\hat{P}(T=x\mid R^{\mathrm{obs}},Z,W,S)\hat{P}(C>x\mid R^{\mathrm{obs}},Z,W,S)\right\}^{\delta}$$
$$\times\left\{\hat{P}(T>x\mid R^{\mathrm{obs}},Z,W,S)\hat{P}(C=x\mid R^{\mathrm{obs}},Z,W,S)\right\}^{1-\delta}, \quad (8)$$

and it is further normalized so that

$$\sum_{i=1}^{m_1}\hat{P}(X=x_i,\Delta=\delta_i\mid R^{\mathrm{obs}}=0,Z=z,W=w,S=1)=1 \text{ if } r^{\mathrm{obs}}=0,$$

$$\sum_{i=m+1}^{n}\hat{P}(X=x_i,\Delta=\delta_i\mid R^{\mathrm{obs}}=1,Z=z,W=w)=1 \text{ if } r^{\mathrm{obs}}=1. \quad (9)$$

We note that in both approaches, the working models for $R^{\mathrm{obs}}=0$ and $R^{\mathrm{obs}}=1$ are variationally independent; i.e., no parameter is shared across working models. A more detailed exposition of how each term in (6) is computed is given in Appendix D.

We then extend the working distribution $\hat{F}$ to $F(\alpha)$, a class of working distributions, by adding a 1-dimensional parameter α to $\hat{P}(X,\Delta\mid R^{\mathrm{obs}},Z,W,S)$ while leaving the other factors of $P_{\hat{F}}$ in (6) unchanged, as follows. If $\hat{P}(X,\Delta\mid R^{\mathrm{obs}},Z,W,S)$ is obtained from Cox

13

models (i.e., approach (i) above), the extended probability parameterized by α is:

$$\hat{P}(X = x, \Delta = \delta \mid R^{\text{obs}} = r^{\text{obs}}, Z, W, S; \alpha)$$
$$= \max[0, \hat{P}(X = x, \Delta = \delta \mid R^{\text{obs}} = r^{\text{obs}}, Z, W, S)\{1 + \alpha x / c_{\max}^{(r^{\text{obs}})}\}], \qquad (10)$$

where $c_{\max}^{(0)} = \max\{c_i : 1 \le i \le m\}$ and $c_{\max}^{(1)} = \max\{c_i : m+1 \le i \le n\}$. For $\alpha > 0$, this extension increases the probability of (x, δ) pairs with larger x values; for $\alpha < 0$, this extension increases the probability of (x, δ) pairs with smaller x values; when $\alpha = 0$, this extension has no effect. The probability calculated by (10) is then normalized to ensure (9).

If $\hat{P}(X, \Delta \mid R^{\text{obs}}, Z, W, S)$ is obtained from log-normal regressions (i.e., approach (ii) above), suppose the fitted regression has mean (after log transformation)

$$\hat{E}\{\log(T) \mid R^{\text{obs}} = 0, Z, W, S = 1\} = \hat{\beta}_0 + Z\hat{\beta}_1 + W\hat{\beta}_2,$$
$$\hat{E}\{\log(T) \mid R^{\text{obs}} = 1, Z, W\} = \hat{\gamma}_0 + Z\hat{\gamma}_1 + W\hat{\gamma}_2.$$

The free parameter α is added to the intercepts $\hat{\beta}_0$ and $\hat{\gamma}_0$, and the extended probability $\hat{P}(X, \Delta \mid R^{\text{obs}}, Z, W, S; \alpha)$ is then computed using (8). For $\alpha > 0$, this extension increases the probability for larger T values; for $\alpha < 0$, this extension increases the probability for smaller T values; when $\alpha = 0$, this extension has no effect. Note that the log-normal regression fits of C is unchanged.

3.4 Remarks on the choice of Ω and $\hat{F}(\alpha)$

We comment on our particular choice of the discretized support Ω and the working distributions $\hat{F}(\alpha)$, and discuss other choices that we have considered.

A necessary condition for Ω is to include all the observed data points to make the perturbation (i.e., $(1 - \epsilon)\hat{F}(\alpha) + \epsilon \mathcal{L}_{O_i}$) easy to compute. This is not sufficient, though. When we attempted to set $\Omega = \{O_1, \ldots, O_n\}$, even though the perturbation is straightforward, the support turned out to be not rich enough to allow the computation of the distribution components in (1) and (2) (especially the distributions involving (X, Δ)). Therefore, we define Ω to be the Cartesian product of unique (z, w) pairs and unique (s, x, δ) triples from the observed data, which allows computation of both the perturbation and the estimand. Note that instead of constructing an overall Cartesian product space, the discretized support is defined within $R^{\text{obs}} = 1$ and $R^{\text{obs}} = 0$, because $\text{pr}(R^{\text{obs}} = 1)$ can be reliably estimated by the

14

empirical frequency and we want the working distribution on Ω to accommodate this. We did not include c_i in constructing Ω, because as shown in (3) its distribution is not needed to identify the parameter of interest.

The choice of an appropriate Ω may depend on the dimension of (Z,W). Suppose, for example, that (Z,W) can only take a few categorical values, then $\Omega' = \{O_1,\ldots,O_n\}$ may be rich enough to allow computation of the distribution components in (1) and (2) when the sample size is relatively large compared to the number of (Z,W)-categories. However, when (Z,W) is multi-dimensional or has continuous variables, a richer discretized support such as the Cartesian product used in Section 3.2 is needed. It is an open question regarding the exact requirements for the discretized support Ω, and how such requirements relate to the dimensionality of the variables involved and to the sample size of the data.

We described in Section 3.3 two approaches to fit the working distribution $\hat{F}$ on Ω, one based on Cox proportional hazards models and the other based on log-normal regressions. Extending $\hat{F}$ to $\hat{F}(\alpha)$ is different for the two approaches: for Cox model-based $\hat{F}$, the free parameter α is added onto the probability distribution of (X,Δ) directly; for log-normal regression-based $\hat{F}$, the free parameter α is added as an intercept to the regression fits for T. A necessary condition for a valid extension $\hat{F}(\alpha)$ is that (5) explores values around 0 as α varies. Other attempts, such as adding the free parameter α to a regression coefficient in the Cox model fits of T and C for Cox model-based $\hat{F}$, have failed due to the fact that (5) does not equal zero for any α value.

We note that the working independence assumption $T \perp C \mid R^{\mathrm{obs}}, Z, W, S$ and other modeling assumptions used in calculating $\hat{P}(X,\Delta \mid R^{\mathrm{obs}}, Z, W, S)$ (proportional hazards assumption for Cox model-based $\hat{F}$, parametric form for log-normal regression-based $\hat{F}$) may not and is not expected to hold, and they are just a convenient way to construct $\hat{P}(X,\Delta \mid R^{\mathrm{obs}}, Z, W, S)$. In fact, Frangakis and Rubin (2001) showed in a setting without covariates that $T \perp C \mid R^{\mathrm{obs}}$ can fail to hold. In the following simulation studies, we see that the proposed estimators are robust to violation of those working assumptions.

4 Simulation

We evaluate the performance of the proposed deductive estimators and compare it with other existing approaches through simulation studies. In Section 4.1 we describe the generative models used in the simulations. In Section 4.2 the proposed estimators are compared

with three other estimators in terms of consistency and coverage probability of confidence intervals. In Section 4.3 we evaluate the performance of the proposed estimators when an incorrect model for the double-sampling selection is used. In Section 4.4 we assess the impact of extending the working model $\hat{F}$ to $\hat{F}(\alpha)$ on the performance of the proposed estimators.

The simulation results show that the proposed estimators are consistent under possibly incorrect working models for both (X,Δ) and for S. The confidence interval can be anti-conservative for small sample sizes, and it becomes close to the nominal level coverage as sample size gets larger. The simulation also shows that the model extension step (from $\hat{F}$ to $\hat{F}(\alpha)$) contributes to the robustness of the estimators.

We briefly summarize the results of some additional simulation studies, the details of which are not listed in the paper. The proposed estimators are not sensitive to the choice of ϵ as long as ϵ is small enough (recall that ϵ is used in calculating the numerical Gateaux derivative in (5)): we varied ϵ from 1×10^{-4} to 1×10^{-8} given the sample sizes used in the simulation study ($n = 50$, 100, 200, 500), and the performance of the proposed estimators stays effectively the same. In addition to the continuous generative model described below, we used a discrete generative model (where Z,T,C,L are all discrete and each has around 10 levels), and the proposed estimators are consistent with nominal level confidence interval coverage.

4.1 Generative models

We consider two generative models. In the first generative model (GM-1), the baseline covariate, Z, is generated from Uniform$[-2,2]$. The true dropout status, R, is generated as Bernoulli random variables with success probability $(Z+3)/6$. The survival time, T, is generated as 5 times a Weilbull-distributed random variable with scale parameter $\exp(Z)$ and shape parameter 5. The administrative censoring time, C, is generated from Uniform$[0.5,2]$, which mimics a continuous accrual process with a constant accrual rate. The dropout time, L, is generated from Uniform$[T/3,T]$ for those with $R=0$, and is NA for those with $R=1$. This is the only longitudinal measurement; i.e., $W=L$. The observed dropout status, R^{obs}, equals 1 for those with $R=1$, and is defined as $R^{\mathrm{obs}} = \mathbb{1}\{\min(T,C) < L\}$ for those with $R=0$, where $\mathbb{1}(\cdot)$ is the indicator function. L is then set to NA for those with $(R=0, R^{\mathrm{obs}}=1)$. The selection for double-sampling, S, is generated as a Bernoulli random variable with success probability $\mathrm{expit}\{(L+Z+1)/2\}$ for those with $R^{\mathrm{obs}}=0$, where

16

$\mathrm{expit}(y) = \{1 + \exp(-y)\}^{-1}$. The survival data, (X, Δ), is defined by $X = \min(T, C)$ and $\Delta = \mathbb{1}(T \leq C)$ for those with $R^{\mathrm{obs}} = 1$ or $(R^{\mathrm{obs}} = 0, S = 1)$, and is (NA, NA) otherwise. The parameter of interest is $\tau = P(T > 0.7) = 77.1\%$.

In the second generative model (GM-2), Z and R are generated the same way as in GM-1. The survival time T and the administrative censoring time C are generated independently from a log-normal distribution LN(Z, 0.25), where $A \sim \mathrm{LN}(\mu, \sigma^2)$ means that $\log(A)$ follows a normal distribution with mean μ and variance σ^2. The dropout time $L \sim \mathrm{Uniform}[T/3, T]$ for those with $R = 0$ and is NA for those with $R = 1$. $R^{\mathrm{obs}}, X, \Delta$ are calculated the same way as in GM-1, and S is generated as Bernoulli with success probability $\mathrm{expit}\{(L - Z + 1)/2\}$ for those with $R^{\mathrm{obs}} = 0$. The parameter of interest is $\tau = P(T > 0.7) = 58.8\%$.

	GM-1	GM-2
$\tau = P(T > 0.7)$	77.1%	58.8%
$P(R^{\mathrm{obs}} = 0)$	0.43	0.36
$P(S = 1 \mid R^{\mathrm{obs}} = 0)$	0.65	0.73
10-th and 90-th percentiles of $P(S = 1 \mid R^{\mathrm{obs}} = 0, Z, L)$	(0.50, 0.80)	(0.65, 0.81)
$P(\Delta = 1 \mid \Delta \neq \mathsf{NA})$	0.70	0.48
10-th and 90-th percentiles of $X \mid X \neq \mathsf{NA}$	(0.53, 1.18)	(0.15, 3.59)
$\mathrm{corr}(T, C \mid Z, R^{\mathrm{obs}})$	0.02	0.24
$\mathrm{corr}(T, C \mid Z, L, R^{\mathrm{obs}} = 0)$	0.00	0.12

Table 1: Descriptive statistics of the two generative models. corr stands for Pearson's partial correlation. All quantities are calculated from a simulated data set with 1,000,000 sample size.

Table 1 lists the descriptive statistics of the two generative models, which are calculated from a simulated data set with 1,000,000 sample size for each GM. For GM-1 (GM-2), 43% (36%) individuals drop out during the study, and among them 65% (73%) are double-sampled. For the patients with $R^{\mathrm{obs}} = 0$, the 10-th and the 90-th percentiles of the probability of being double-sampled are 0.50 and 0.80 for GM-1 (0.65 and 0.81 for GM-2). Among the patients with observed (X, Δ), 70% are dead on observation ($\Delta = 1$) for GM-1 (48% for GM-2), and the 10-th and the 90-th percentiles of X are 0.53 and 1.18 for GM-1 (0.15 and 3.59 for GM-2). The Pearson's partial correlation is $\mathrm{corr}(T, C \mid Z, R^{\mathrm{obs}}) = 0.02$ for GM-1 (0.24 for GM-2), and $\mathrm{corr}(T, C \mid Z, L, R^{\mathrm{obs}} = 0) = 0.00$ for GM-1 (0.12 for GM-2); these are computed using R package ppcor (Kim, 2015). This implies that although As-

sumption 2 holds for both GMs, independent censoring conditional on Z, R^{obs} or Z, L, R^{obs} does not hold for GM-2 and may be slightly violated for GM-1. In addition, in GM-1 the proportional hazards assumption on $T \mid Z$ holds, and in GM-2 the proportional hazards assumption on $T \mid Z$ does not hold (Bender et al., 2005).

4.2   Simulation on consistency

We evaluate the performance of the proposed estimators in terms of consistency and confidence interval coverage and compare them with other existing methods. We consider the following five estimators:

- DE.Cox: the deductive estimator proposed in Section 3 with Cox model as working model for the distribution of (X, Δ). (Z, L) are included in both the Cox model fits and the model for double-sampling selection S. ϵ is set to be 1×10^{-4}.

- DE.LN: the deductive estimator proposed in Section 3 with log-normal regression as working model for the distribution of (X, Δ). (Z, L) are included in both the log-normal regression and the model for double-sampling selection S. ϵ is set to be 1×10^{-4}.

- PAR: the estimator in An et al. (2014) that is based on parametric assumptions. In this particular implementation we used, $T \mid R = 1, Z$ and $T \mid R = 0, Z, L$ are assumed to follow log-normal distributions. The R code to compute the estimator and the standard error is provided by Dr. Ming-Wen An.

- KM.S: a stratified Kaplan-Meier estimator with stratification variable R^{obs}. In this estimator, a Kaplan-Meier estimator (Kaplan and Meier, 1958) is computed separately for the patients with $(R^{\text{obs}} = 0, S = 1)$ and for the patients with $R^{\text{obs}} = 1$, and a weighted average is taken using weights $\hat{P}(R^{\text{obs}})$. The bias-corrected and accelerated (BCa) bootstrap 95% confidence interval is obtained using R package boot (Canty and Ripley, 2017), with 1000 bootstrap replicates.

- KM.C: a Kaplan-Meier estimator using only complete cases (i.e., data with $R^{\text{obs}} = 1$ or $(R^{\text{obs}} = 0, S = 1)$) without any weighting.

The simulation result is shown in Table 2. Simulation is conducted for sample size $n = 50, 100, 200, 500$, each with 1000 replicates. Both DE.Cox and DE.LN are consistent

GM	n	DE.Cox			DE.LN			PAR			KM.S			KM.C		
		Bias	CP	SD	Bias	CP	SD	Bias	CP	SD	Bias	CP	SD	Bias	CP	SD
1	50	0.6	92.1	6.5	0.0	91.6	6.7	-1.5	89.9	6.0	-0.8	95.6	6.9	-1.4	97.7	6.8
	100	-0.1	94.2	4.6	-0.2	93.7	4.6	-1.9	89.4	4.2	-1.1	95.1	4.9	-1.7	98.6	4.9
	200	0.2	93.2	3.3	0.1	92.7	3.3	-1.9	87.4	3.0	-0.8	93.9	3.4	-1.4	98.8	3.4
	500	0.0	93.9	2.0	0.0	93.6	2.0	-2.1	74.8	1.9	-0.9	93.0	2.2	-1.5	98.1	2.2
2	50	2.3	85.6	9.2	0.4	88.7	11.6	0.2	92.8	7.2	12.7	67.4	7.8	15.0	62.6	7.5
	100	0.3	89.8	6.2	0.7	91.1	7.1	-0.2	94.2	4.7	12.9	36.9	5.3	15.2	37.6	5.0
	200	-0.2	92.1	4.1	0.7	93.4	4.5	-0.2	94.4	3.3	12.9	8.7	3.7	15.2	8.8	3.5
	500	-0.4	92.8	2.5	0.4	93.8	2.7	-0.4	94.7	2.0	12.6	0.1	2.3	14.9	0.1	2.3

* Note: Bias and SD in the table have been multiplied by 100, so the Bias and SD are on the scale of percentage survival probability.

Table 2: Simulation result for estimators DE.Cox, DE.LN, PAR, KM.S aand KM.C. GM: generative model; n: sample size; CP: coverage probability (in %) of 95% confidence interval based on normal approximation; SD: standard deviation. All results are based on 1000 replicates.

under both GM-1 and GM-2; their coverage probability of 95% confidence interval based on normal approximation is anti-conservative for small sample sizes such as $n = 50$ (with actual coverage probability being around 92%), but it becomes close to the nominal level for $n = 500$. PAR is consistent under GM-2 because $T \mid R, Z$ follows log-normal distribution, and it is more efficient (i.e., has smaller standard deviation) than the other estimators because of the parametric assumption; it is inconsistent under GM-1 because the log-normal parametric assumption is violated there. KM.S is possibly inconsistent under GM-1, as the coverage probability starts to decrease as n gets larger; however, this possible inconsistency is small because $\mathrm{corr}(T, C \mid Z, R^{\mathrm{obs}}) = 0.02$, which means that the assumption $T \perp C \mid Z, R^{\mathrm{obs}}$ required for its consistency is only slightly violated. KM.S is severely inconsistent under GM-2, where $\mathrm{corr}(T, C \mid Z, R^{\mathrm{obs}}) = 0.24$. KM.C is severely inconsistent under GM-2; under GM-1, it is possibly inconsistent because the bias does not decrease as n grows, but its confidence interval is conservative and has larger than nominal coverage probability.

The simulation result shows the potential robustness of DE.Cox and DE.LN. Under GM-2, the working independence assumption $T \perp C \mid Z, L, R^{\mathrm{obs}}$ and the proportional hazards assumption for $T \mid Z, L, R^{\mathrm{obs}}$ are both likely violated, yet DE.Cox is consistent with close to nominal confidence interval coverage especially when n is large. Under GM-1, the log-normal distributional assumption $T \mid Z, L, R^{\mathrm{obs}}$ is likely violated, yet DE.LN is consistent

with close to nominal confidence interval coverage especially when n is large.

4.3   Simulation with incorrect double-sampling model for S

We evaluate the performance of DE.Cox and DE.LN when an incorrect model for the double-sampling selection, $P(S=1 \mid R^{\mathrm{obs}}=0,Z,L)$, is used. In particular, let DE.Cox.WrongS and DE.LN.WrongS denote the estimators DE.Cox and DE.LN, respectively, when the model $\hat{P}(S=1 \mid R^{\mathrm{obs}}=0,Z,L)$ only includes an intercept (i.e., regressors Z,L are omitted). The simulation result is shown in Table 3. For both estimators, the performances under correct model for S and under incorrect model for S are comparable under both GMs, indicating that the proposed estimators are likely robust to incorrect models of the double-sampling selection.

		DE.Cox			DE.Cox.WrongS			DE.LN			DE.LN.WrongS		
GM	n	Bias	CP	SD	Bias	CP	SD	Bias	CP	SD	Bias	CP	SD
	50	0.6	92.1	6.5	0.4	93.3	6.5	0.0	91.6	6.7	0.1	92.3	6.7
1	100	-0.1	94.2	4.6	0.0	94.8	4.5	-0.2	93.7	4.6	-0.1	94.0	4.5
	200	0.2	93.2	3.3	0.3	94.2	3.3	0.1	92.7	3.3	0.3	93.1	3.3
	500	0.0	93.9	2.0	0.2	94.5	2.0	0.0	93.6	2.0	0.1	93.9	2.0
	50	2.3	85.6	9.2	2.3	85.9	9.3	0.4	88.7	11.6	0.4	88.8	11.6
2	100	0.3	89.8	6.2	0.3	89.7	6.2	0.7	91.1	7.1	0.7	90.9	7.1
	200	-0.2	92.1	4.1	-0.2	92.2	4.1	0.7	93.4	4.5	0.7	93.1	4.5
	500	-0.4	92.8	2.5	-0.4	92.8	2.5	0.4	93.8	2.7	0.4	93.8	2.7

* Note: Bias and SD in the table have been multiplied by 100, so the Bias and SD are on the scale of percentage survival probability.

Table 3: Simulation result for estimators DE.Cox and DE.LN, and when they are using incorrect model for $P(S=1 \mid R^{\mathrm{obs}}=0,Z,W)$ (columns with WrongS). GM: generative model; n: sample size; CP: coverage probability (in %) of 95% confidence interval based on normal approximation; SD: standard deviation. All results are based on 1000 replicates. The columns for DE.Cox and DE.LN are the same as those in Table 2, and are included here for comparison.

4.4 Simulation on the impact of α

We evaluate the impact of the free parameter α on the performance of the proposed estimators. In particular, we consider two additional estimators: DE.Cox($\alpha = 0$) and DE.LN($\alpha = 0$). DE.Cox($\alpha = 0$) uses the same working models as DE.Cox, but it did not solve (5) for $\hat{\alpha}$ and output as estimator $\tau\{\hat{F}(\hat{\alpha})\}$; instead, it sets $\alpha = 0$ and outputs as estimator $\tau\{\hat{F}(0)\}$. DE.LN($\alpha = 0$) uses the same working models as DE.LN, but it sets $\alpha = 0$ and outputs as estimator $\tau\{\hat{F}(0)\}$. Recall that for DE.Cox and DE.LN, $\alpha = 0$ means that no extension is added to the fitted working distribution $\hat{F}$.

		DE.Cox			DE.Cox($\alpha = 0$)			DE.LN			DE.LN($\alpha = 0$)		
GM	n	Bias	CP	SD	Bias	CP	SD	Bias	CP	SD	Bias	CP	SD
	50	0.6	92.1	6.5	1.1	92.8	6.3	0.0	91.6	6.7	-2.9	83.7	8.3
1	100	-0.1	94.2	4.6	0.3	94.2	4.5	-0.2	93.7	4.6	-3.3	82.1	5.9
	200	0.2	93.2	3.3	0.3	93.4	3.3	0.1	92.7	3.3	-3.1	76.9	4.1
	500	0.0	93.9	2.0	0.0	94.3	2.0	0.0	93.6	2.0	-3.5	55.7	2.6
	50	2.3	85.6	9.2	5.2	82.7	9.5	0.4	88.7	11.6	3.5	85.8	10.4
2	100	0.3	89.8	6.2	1.5	89.4	6.1	0.7	91.1	7.1	2.6	89.8	6.9
	200	-0.2	92.1	4.1	0.3	93.5	4.0	0.7	93.4	4.5	2.6	89.9	4.7
	500	-0.4	92.8	2.5	-0.5	92.9	2.5	0.4	93.8	2.7	2.1	84.4	2.9

* Note: Bias and SD in the table have been multiplied by 100, so the Bias and SD are on the scale of percentage survival probability.

Table 4: Simulation result for estimators DE.Cox, DE.Cox($\alpha = 0$), DE.LN, and DE.LN($\alpha = 0$). GM: generative model; n: sample size; CP: coverage probability (in %) of 95% confidence interval based on normal approximation; SD: standard deviation. All results are based on 1000 replicates. The columns for DE.Cox and DE.LN are the same as those in Table 2, and are included here for comparison.

The simulation result is shown in Table 4. DE.Cox and DE.Cox($\alpha = 0$) have very similar performance under both GMs. On the other hand, DE.LN and DE.LN($\alpha = 0$) have very different performance: DE.LN is consistent, whereas DE.LN($\alpha = 0$) is severely inconsistent under both GMs. This shows the robustness brought by the $\hat{\alpha}$ that solves (5).

5 Application to PEPFAR

We apply the proposed method to estimating the mortality rate using the data set from a double-sampling design component of the President's Emergency Plan for Aids Relief (PEPFAR), and HIV monitoring and treatment program in East Africa evaluating the antiretroviral treatment (ART) for HIV-infected people (Geng et al., 2015). The data set consists of 1,773 HIV-infected adults from Morogoro, Tanzania, who started ART after entering the study. 673 patients (38%) dropped out during the study. Among the dropouts, 91 (14% of the dropouts) got double-sampled. We use baseline age and pre-treatment CD4 value as Z, and the loss to follow-up time L and the CD4 value measured at the last visit before dropout as W. The first column in Figure 1 gives the number of patients of each type; for the types that are not distinguishable from the observed data (such as (b1) and (b2)), a single number is reported.

Estimated mortality rate. We use the estimators proposed in Section 3 and two other estimators considered in the simulation study (Section 4.2) to estimate the mortality rate $P(T \leq t)$ for $0 \leq t \leq 2$ using the PEPFAR data. Figure 2 shows the result, with solid curves representing the estimates and dashed curves representing the 95% pointwise confidence intervals. The deductive estimator using Cox proportional hazards model (DE.Cox, in black) and the stratified Kaplan-Meier estimator (KM.S, in blue) give very similar mortality rate estimates, although the confidence interval of KM.S is wider. The deductive estimator using log-normal regression (DE.LN, in yellow) gives slightly higher mortality rate estimates, with confidence interval width comparable to DE.Cox. The complete-case Kaplan-Meier estimator (KM.C, in green) gives much lower estimates than the other three; it is known to be biased for double-sampling designs (Frangakis and Rubin, 2001), and we include it for reference. The estimates and the confidence intervals for $t = 0.5$, 1, 1.5, 2 are listed in the first four rows (with $\gamma = \infty$) in Table 5.

Impact of the double-sampling selection criteria on the estimated mortality rate. In practice, to double-sample a dropout may require a considerable amount of effort and resource; therefore, it may be more feasible to double-sample relatively recent dropouts (An et al., 2014). We evaluate the impact of such restrictions on double-sampling selection on the estimated mortality, by creating pseudo data sets based on the PEPFAR data, as follows. Suppose an investigator decides that only the *recent* dropouts have positive probability to be double-sampled, where a dropout is *recent* if his/her dropout time is within the past
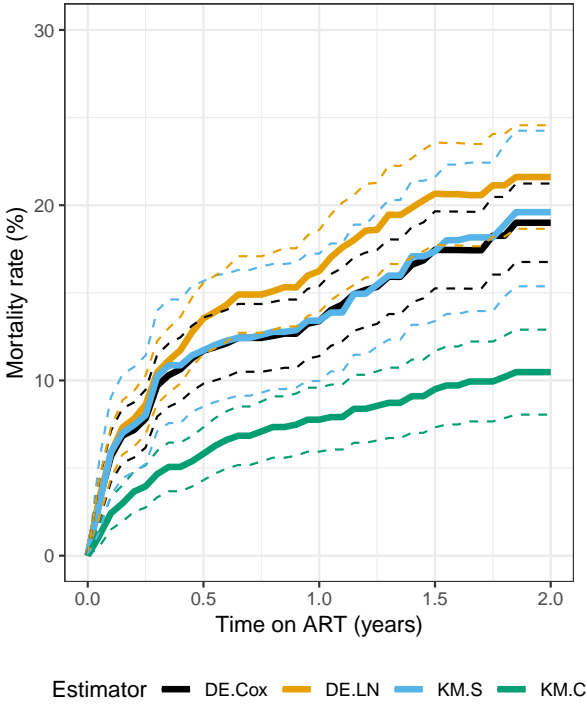
Figure 2: Estimated mortality rate (solid line) with 95% pointwise confidence interval (dashed line) using all patients data from PEPFAR with four estimators: DE.Cox, DE.LN, KM.S, KM.C.
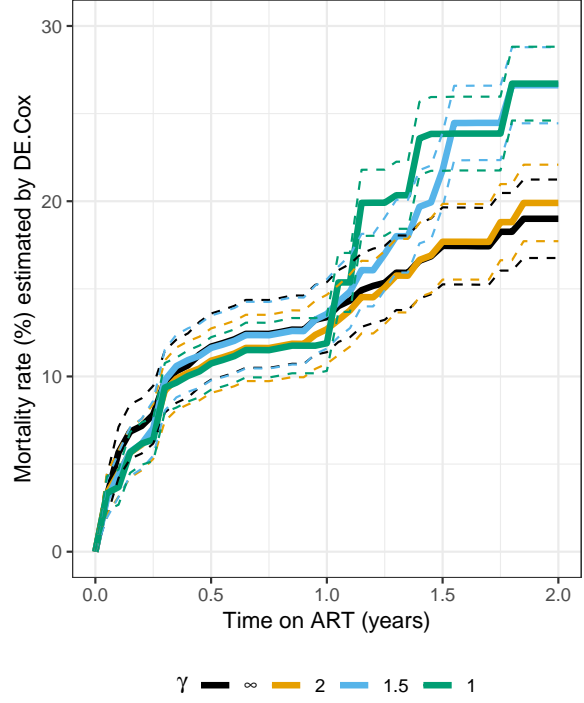
Figure 3: Estimated mortality rate (solid line) with 95% pointwise confidence interval (dashed line) estimated by DE.Cox, using PEPFAR patients data with various γ restrictions: all double-samples included ($\gamma = \infty$), past 2-year dropouts included ($\gamma = 2$), past 1.5-year dropouts included ($\gamma = 1.5$), past 1-year dropouts included ($\gamma = 1$).

γ	t	DE.Cox mort	DE.Cox CI	DE.LN mort	DE.LN CI	KM.S mort	KM.S CI	KM.C mort	KM.C CI
	0.5	11.7	(9.8, 13.6)	13.5	(11.5, 15.6)	11.7	(8.5, 15.7)	5.8	(4.3, 7.3)
∞	1.0	13.4	(11.4, 15.4)	16.2	(13.9, 18.6)	13.4	(10.0, 17.2)	7.8	(5.9, 9.6)
	1.5	17.5	(15.3, 19.7)	20.7	(17.7, 23.6)	17.4	(13.4, 21.6)	9.5	(7.3, 11.7)
	2.0	19.0	(16.8, 21.2)	21.6	(18.7, 24.6)	19.6	(15.4, 24.2)	10.5	(8.1, 12.9)
	0.5	10.9	(9.1, 12.7)	13.4	(11.4, 15.4)	10.4	(7.3, 14.7)	5.3	(3.9, 6.8)
2	1.0	12.7	(10.7, 14.7)	16.4	(14.0, 18.7)	12.2	(8.7, 16.2)	7.3	(5.5, 9.1)
	1.5	17.7	(15.5, 19.8)	21.3	(18.3, 24.3)	16.9	(12.8, 21.8)	9.1	(6.9, 11.2)
	2.0	19.9	(17.7, 22.1)	22.3	(19.3, 25.3)	19.8	(15.3, 25.5)	10.1	(7.7, 12.5)
	0.5	11.6	(9.8, 13.5)	15.3	(13.2, 17.4)	11.7	(7.7, 16.3)	5.3	(3.9, 6.8)
1.5	1.0	13.6	(11.6, 15.5)	18.8	(16.3, 21.3)	13.6	(9.5, 18.3)	7.3	(5.5, 9.1)
	1.5	21.8	(19.7, 23.9)	25.0	(21.9, 28.1)	19.9	(14.9, 25.6)	9.2	(7.0, 11.3)
	2.0	26.6	(24.4, 28.8)	27.5	(24.4, 30.5)	24.4	(18.3, 31.0)	10.2	(7.7, 12.7)
	0.5	10.7	(9.2, 12.2)	16.6	(14.6, 18.6)	11.7	(6.7, 17.8)	4.8	(3.4, 6.2)
1	1.0	11.9	(10.3, 13.5)	19.6	(17.3, 22.0)	13.0	(7.9, 19.1)	6.7	(4.9, 8.4)
	1.5	23.8	(21.7, 25.9)	23.9	(21.1, 26.7)	18.6	(11.6, 26.8)	7.8	(5.8, 9.9)
	2.0	26.7	(24.6, 28.8)	25.7	(22.9, 28.4)	21.1	(13.7, 29.6)	8.4	(6.2, 10.5)

Table 5: Estimated t-year mortality rate (mort, in %) and 95% confidence intervals (CI) using estimators DE.Cox, DE.LN, KM.S, and KM.C. $\gamma = \infty$ means that the estimates are calculated from the original PEPFAR data. $\gamma = 2$, 1.5, or 1 means that the estimates are calculated from the pseudo data set where only the past 2-, 1.5-, or 1-year dropouts are possible for double-sampling.

γ years; in other words, if $C_i - L_i \leq \gamma$. A pseudo data set with regard to γ is created by setting $S_i = 0$ and $(X_i, \Delta_i) = (\mathsf{NA}, \mathsf{NA})$ for those in the PEPFAR data with $C_i - L_i > \gamma$. Figure 3 gives the estimated morality rates (in solid curves) and the 95% pointwise confidence intervals (in dashed curves) for the pseudo data sets with different γ, using the deductive estimator with Cox proportional hazards model (DE.Cox). Four γ values are considered: $\gamma = \infty$ (no restrictions on double-sampling selection, which corresponds to 91 double-samples in the original PEPFAR data), and $\gamma = 2$, 1.5, 1 (only the past 2-, 1.5-, or 1-year dropouts are possible for double-sampling, which corresponds to 77, 62, or 33 double-samples in the pseudo data set). By definition, the curve for $\gamma = \infty$ (black curve in Figure 3) is the same as the black curve in Figure 2. We see that for DE.Cox, the curve for $\gamma = 2$ (in yellow) is similar to that for $\gamma = \infty$ throughout $0 \leq t \leq 2$; the curve for $\gamma = 1.5$ (in blue) starts to diverge from that for $\gamma = \infty$ at around $t = 1.5$; the curve for $\gamma = 1$ (in green) starts to diverge from that for $\gamma = \infty$ at around $t = 1$. Figures of various γ values for the other three estimators (DE.LN, KM.S, KM.C) are included in Appendix E. Although the exact same pattern (the curve for $\gamma = \gamma_0$ diverging from that for $\gamma = \infty$ at around $t = \gamma_0$) is not observed for the other estimators, the general conclusion is that the more restrictive the selection criteria is (i.e., smaller γ), the more different the estimate is from that when there is no restriction. This is because for a given $\gamma < \infty$, no double-samples would have been enrolled in the study for longer than γ years, and therefore little information is available regarding the beyond-γ survival probability for the dropouts and such estimate will be largely based on extrapolation. The estimates and the confidence intervals for $t = 0.5$, 1, 1.5, 2 using pseudo data sets with $\gamma = \infty$, 2, 1.5, 1 are listed in Table 5.

6   Discussion

We proposed a deductive method to produce semiparametric estimators for estimating survival probability in the double-sampling design. The method generalizes the approach in Frangakis et al. (2015) by incorporating the discretized support structure and is easily computerizable. Two implementations of the method are described: one using Cox models as working models, and the other using log-normal regressions as working models. We applied the method to estimating mortality rate using data from a double-sampling component at a site of the PEPFAR program, and evaluated the impact of double-sampling selection criteria on the estimated mortality rate.

The simulation results show that the confidence interval calculated from normal approximation and the influence function-based standard error, $n^{-1}\big[\sum_i \mathsf{Gateaux}\{O_i, \hat{F}(\hat{\alpha}), \epsilon\}^2\big]^{1/2}$, can be anti-conservative for the deductive estimators with lower than nominal level coverage probability, especially for small sample size such as $n = 50$ or 100. A potential alternative is to use bootstrap to obtain the confidence intervals; however, we were not able to numerically validate this due to the computational cost of bootstrapping the deductive estimator.

It would be interesting to compare the proposed estimators with estimators derived if one had used the explicit form of the EIF. While Robins and Rotnitzky (2001) describe briefly the conditions that the EIF satisfies through a set of equations, we have found that by solving those equations by hand, we had high risk of introducing possible errors. We are also unaware of any work that gives a closed form expression of the EIF.

In the literature, discussions on robust estimators have been partly based on characterizing robust estimating functions; see, for example, Robins et al. (2000) and Robins and Rotnitzky (2001). Although our estimators are obtained by (numerically) solving the EIF equation, it is more difficult to find analytically all the conditions for which the estimators are consistent, precisely because the focus is on problems in which the EIF is difficult to derive analytically. Perhaps, therefore, a supplemental numerical method may exist that can also characterize more intuitively these conditions.

R code for the simulation and the data analysis in the paper can be downloaded at https://github.com/tqian/dce_ds.

Acknowledgments

References

An MW, Frangakis CE, Musick BS, Yiannoutsos CT (2009) The need for double-sampling designs in survival studies: An application to monitor PEPFAR. Biometrics 65(1):301–306

An MW, Frangakis CE, Yiannoutsos CT (2014) Choosing profile double-sampling designs for survival estimation with application to President's Emergency Plan for Aids Relief evaluation. Statistics in medicine 33(12):2017–2029

Baker SG, Wax Y, Patterson BH (1993) Regression analysis of grouped survival data: informative censoring and double sampling. Biometrics 49(2):379–389

Bender R, Augustin T, Blettner M (2005) Generating survival times to simulate cox proportional hazards models. Statistics in medicine 24(11):1713–1723

Bickel PJ, Klaassen CA, Ritov Y, Wellner JA (1993) Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins University Press, Baltimore

Breslow N (1974) Covariance analysis of censored survival data. Biometrics pp 89–99

Canty A, Ripley BD (2017) boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20

Chamberlain G (1987) Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics 34(3):305–334

Cochran WG (2007) Sampling techniques. John Wiley & Sons

Cox DR (1972) Regression models and life tables. Journal of the Royal Statistical Society, Ser B 34(2):187–220

Frangakis CE, Rubin DB (2001) Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. Biometrics 57(2):333–342

Frangakis CE, Qian T, Wu Z, Diaz I (2015) Deductive derivation and Turing-computerization of semiparametric efficient estimation. Biometrics 71(4):867–874

Geng EH, Odeny TA, Lyamuya RE, Nakiwogga-Muwanga A, Diero L, Bwana M, Muyindike W, Braitstein P, Somi GR, Kambugu A, et al. (2015) Estimation of mortality among HIV-infected people on antiretroviral treatment in east Africa: a sampling based approach in an observational, multisite, cohort study. The Lancet HIV 2(3):e107–e116

Glynn RJ, Laird NM, Rubin DB (1993) Multiple imputation in mixture models for non-ignorable nonresponse with follow-ups. Journal of the American Statistical Association 88(423):984–993

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. Journal of the American statistical association 53(282):457–481

Kim S (2015) ppcor: Partial and Semi-Partial (Part) Correlation. URL https://CRAN.R-project.org/package=ppcor, r package version 1.1

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1):13–22

Newey WK (1990) Semiparametric efficiency bounds. Journal of applied econometrics 5(2):99–135

Robins J, Rotnitzky A (2001) Comments. Statistica Sinica 11(4):920–936

Robins J, Rotnitzky A, Bonetti M (2001) Discussion of the Frangakis and Rubin article. Biometrics 57(2):343–347

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89(427):846–866

Robins JM, Rotnitzky A, van der Laan M (2000) Comment. Journal of the American Statistical Association 95(450):477–482, DOI 10.1080/01621459.2000.10474224

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Therneau TM (2015) A Package for Survival Analysis in S. URL https://CRAN.R-project.org/package=survival, version 2.38

Tsiatis AA (2007) Semiparametric Inference and Missing Data. Springer

Turing AM (1937) On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society 2(1):230–265

van der Vaart AW (2000) Asymptotic Statistics. Cambridge University Press

Zhang M, Tsiatis AA, Davidian M (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. Biometrics 64(3):707–715

Appendix

A   Robustness of the deductive estimator

Suppose the working model assumes the true distribution F_0 belongs in some set $\mathcal{F}$. Then the estimator, say $\hat{\tau}$, that solves the nonparametric EIF within the working model will, in the limit, be

$$\tau(F^*), \text{ for } F^* \in \mathcal{F} \text{ that solves } E_0\phi(D_i, F^*) = 0, \tag{A.1}$$

where E_0 denotes the expectation under F_0. Therefore, by denoting $\bar{\phi}(F_0, F) := E_0\phi(D_i, F)$, and assuming sufficient smoothness of the distributions, the estimator $\hat{\tau}$ will converge to the true value $\tau(F_0)$ under the following joint conditions:

$$\begin{cases} \tau(F^*) = \tau(F_0) & \text{(providing correct estimand)} \\ F^* \text{ solves } \bar{\phi}(F_0, F^*) = 0 & \text{(fitting the model } \{F\} \text{ using } \phi) \end{cases} \tag{A.2}$$

The analytic form of above expressions may not be easily accessible when the form of the EIF is not. Suppose however, a computational method can easily determine just the "zeros" of the expressions, that is, given any F_0, the features of F^* that are restricted in order to satisfy conditions (A.2). With such a method, coupled with the method of deductive estimation, the researcher can focus efforts to clarify and model especially well those restricted features, as this would provide approximate consistency of the estimator.

B   Identification of $P(T > t)$

Suppose $\mathrm{surv}_t(\cdot)$ is a function that takes an arbitrary distribution $\mathcal{L}$ of (X, Δ) from independent survival and censoring times and returns the survival probability beyond t, $P(T > t)$ (this function is the probability limit of the Kaplan-Meier estimator (Kaplan and Meier, 1958)). By Assumption 2 ($T \perp C \mid Z$), we have $P(T > t \mid Z = z) = \mathrm{surv}_t\{\mathcal{L}(X, \Delta \mid z)\}$, where $\mathcal{L}(X, \Delta \mid z)$ denotes the conditional distribution of (X, Δ) given $Z = z$. Thus,

$$P(T > t) = E_Z[\mathrm{surv}_t\{\mathcal{L}(X, \Delta \mid Z)\}].$$

Let $p_{X,\Delta|Z}(x,\delta\,|\,z)$ denote the probability density function of $\mathcal{L}(X,\Delta\,|\,Z)$. We have

$$p_{X,\Delta|Z}(x,\delta\,|\,z) = p_{X,\Delta,R^{\mathrm{obs}}|Z}(x,\delta,1\,|\,z) + p_{X,\Delta,R^{\mathrm{obs}}|Z}(x,\delta,0\,|\,z). \tag{B.1}$$

By basic probability manipulation, we have

$$\begin{aligned} p_{X,\Delta,R^{\mathrm{obs}}|Z}(x,\delta,1\,|\,z) &= p_{X,\Delta,R^{\mathrm{obs}},Z}(x,\delta,1,z)/p_Z(z) \\ &= p_{R^{\mathrm{obs}}}(1)p_{Z|R^{\mathrm{obs}}}(z\,|\,1)p_{X,\Delta|R^{\mathrm{obs}},Z}(x,\delta\,|\,1,z)/p_Z(z), \end{aligned} \tag{B.2}$$

and

$$\begin{aligned} &p_{X,\Delta,R^{\mathrm{obs}}|Z}(x,\delta,0\,|\,z) \\ &= p_{X,\Delta,R^{\mathrm{obs}},Z}(x,\delta,0,z)/p_Z(z) \\ &= \int p_{X,\Delta,R^{\mathrm{obs}},Z,W}(x,\delta,0,z,w)\mu(dw)/p_Z(z) \\ &= \int p_{R^{\mathrm{obs}}}(0)p_{Z,W|R^{\mathrm{obs}}}(z,w\,|\,0)p_{X,\Delta|R^{\mathrm{obs}},Z,W,S}(x,\delta\,|\,0,z,w,1)\mu(dw)/p_Z(z), \end{aligned} \tag{B.3}$$

where μ is the dominating measure, and (B.3) follows from Assumption 1; i.e.,

$$p_{X,\Delta|R^{\mathrm{obs}},Z,W}(x,\delta\,|\,0,z,w) = p_{X,\Delta|R^{\mathrm{obs}},Z,W,S}(x,\delta\,|\,0,z,w,1).$$

Plugging (B.2) and (B.3) into (B.1), we have that $p_{X,\Delta|Z}(x,\delta\,|\,z)$, and hence $P(T>t)$, is identifiable from the observed data distribution components (1) and (2).

C Computing (1) and (2) from an arbitrary distribution on Ω

Suppose G is an arbitrary probability distribution on $\Omega = \Omega_0 \cup \Omega_1$. Because Ω is discrete and so is G, we will use $P(\cdot)$ instead $\mathrm{pr}(\cdot)$. For an arbitrary point $o \in \Omega$, we use $o(z)$ to denote the entry of o corresponding to Z, and similarly for $o(x)$, $o(\delta)$, etc. We define $0/0 = 0$. The distribution components (1) and (2) can be computed as follows.

$$P(R^{\text{obs}} = 1) = \sum_{o \in \Omega_1} G(o),$$

$$P(Z = z^* \mid R^{\text{obs}} = 1) = \frac{\sum_{o \in \Omega_1} G(o) \mathbb{1}\{o(z) = z^*\}}{\sum_{o \in \Omega_1} G(o)},$$

$$P(X = x^*, \Delta = \delta^* \mid Z = z^*, R^{\text{obs}} = 1) = \frac{\sum_{o \in \Omega_1} G(o) \mathbb{1}\{o(x) = x^*, o(\delta) = \delta^*, o(z) = z^*\}}{\sum_{o \in \Omega_1} G(o) \mathbb{1}\{o(z) = z^*\}},$$

$$P(R^{\text{obs}} = 0) = 1 - P(R^{\text{obs}} = 1),$$

$$P(Z = z^*, W = w^* \mid R^{\text{obs}} = 0) = \frac{\sum_{o \in \Omega_0} G(o) \mathbb{1}\{o(z) = z^*, o(w) = w^*\}}{\sum_{o \in \Omega_0} G(o)},$$

and

$$P(X = x^*, \Delta = \delta^* \mid Z = z^*, W = w^*, R^{\text{obs}} = 0, S = 1)$$
$$= \frac{\sum_{o \in \Omega_0} G(o) \mathbb{1}\{o(x) = x^*, o(\delta) = \delta^*, o(z) = z^*, o(w) = w^*, s = 1\}}{\sum_{o \in \Omega_1} G(o) \mathbb{1}\{o(z) = z^*, o(w) = w^*, s = 1\}}.$$

D   Details about the computation of terms in (6)

Recall that the observed data for patient i is denoted by $(r_i^{\text{obs}}, z_i, w_i, s_i, x_i, \delta_i)$, and we assume that the data is sorted so that $r_i^{\text{obs}} = 0$ for $1 \leq i \leq m$, $r_i^{\text{obs}} = 1$ for $m+1 \leq i \leq n$, $s_i = 1$ for $1 \leq i \leq m_1$, and $s_i = 0$ for $m_1 + 1 \leq i \leq n$, and that $x_1 < x_2 < \dots < x_{m_1}$ and $x_{m+1} < x_{m+2} < \dots < x_n$. This last assumption that no ties in x values is for notation simplicity in the following exposition; our estimation program allows ties in x values.

Let $o = (r^{\text{obs}}, z, w, s, x, \delta)$ denote an arbitrary point in Ω (see Section 3.2 for the definition of Ω), and let $P_{\hat{F}}(o)$ denote the probability of o under the distribution $\hat{F}$. We rewrite (6) here for reference:

$$P_{\hat{F}}(r^{\text{obs}}, z, w, s, x, \delta) = \hat{P}(R^{\text{obs}} = r^{\text{obs}}) \hat{P}(Z = z, W = w \mid R^{\text{obs}} = r^{\text{obs}})$$
$$\times \hat{P}(S = s \mid R^{\text{obs}} = r^{\text{obs}}, Z = z, W = w)$$
$$\times \hat{P}(X = x, \Delta = \delta \mid R^{\text{obs}} = r^{\text{obs}}, Z = z, W = w, S = s).$$

Each of the terms on the right hand side is computed as follows:

- $\hat{P}(R^{\text{obs}} = r^{\text{obs}})$: the probability of being an observed dropout. This is fitted by the

empirical frequency: $\hat{P}(R^{\mathrm{obs}} = r^{\mathrm{obs}}) = n^{-1}\sum_{i=1}^{n}\mathbb{1}(r_i^{\mathrm{obs}} = r^{\mathrm{obs}})$.

- $\hat{P}(Z = z, W = w \mid R^{\mathrm{obs}} = r^{\mathrm{obs}})$. This is fitted by the empirical frequency:

$$\hat{P}(Z = z, W = w \mid R^{\mathrm{obs}} = r^{\mathrm{obs}}) = \frac{\sum_{i=1}^{n}\mathbb{1}(r_i^{\mathrm{obs}} = r^{\mathrm{obs}}, z_i = z, w_i = w)}{\sum_{i=1}^{n}\mathbb{1}(r_i^{\mathrm{obs}} = r^{\mathrm{obs}})}.$$

- $\hat{P}(S = 1 \mid R^{\mathrm{obs}} = 0, Z = z, W = w)$. This is fitted by logistic regression using data from patients with $r_i^{\mathrm{obs}} = 0$ and with Z, W being regressors. For $R^{\mathrm{obs}} = 1$, S is always equal to 0 by the construction of Ω.

- $\hat{P}(X = x, \Delta = \delta \mid R^{\mathrm{obs}} = r^{\mathrm{obs}}, Z = z, W = w, S = s)$. For $(r^{\mathrm{obs}} = 0, s = 0)$, by the construction of Ω we have $(x, \delta) = (\mathsf{NA}, \mathsf{NA})$, and hence the corresponding probability equals 1. For $(r^{\mathrm{obs}} = 0, s = 1)$ or $r^{\mathrm{obs}} = 1$, this probability is fitted using the likelihood arising from one of two working models: Cox proportional hazards model (Cox, 1972) or log-normal regression. In both working models, the working assumption of conditional independent censoring $(T \perp C \mid R^{\mathrm{obs}}, Z, W, S)$ is used.

    - When Cox proportional hazards model is used as the working model, we fit four separate Cox models without sharing parameters across models:

$$\begin{aligned} T \sim Z, W \mid R^{\mathrm{obs}} = 0, S = 1; &\qquad C \sim Z, W \mid R^{\mathrm{obs}} = 0, S = 1; \\ T \sim Z, W \mid R^{\mathrm{obs}} = 1; &\qquad C \sim Z, W \mid R^{\mathrm{obs}} = 1. \end{aligned} \qquad \text{(D.1)}$$

    For example, $T \sim Z, W \mid R^{\mathrm{obs}} = 0, S = 1$ means that we fit a Cox model with T being the failure time (hence C being the censoring time), with Z, W as regressors, and using data from patients with $R^{\mathrm{obs}} = 0, S = 1$; $C \sim Z, W \mid R^{\mathrm{obs}} = 1$ means that we fit a Cox model with C being the failure time (hence T being the censoring time), with Z, W as regressors, and using data from patients with $R^{\mathrm{obs}} = 1$. For each model the baseline hazard is estimated using Breslow's method (Breslow, 1974). This gives the following survival probabilities:

$$\begin{aligned} S_T^{(0)}(x; z, w) &:= \hat{P}(T > x \mid Z = z, W = w, R^{\mathrm{obs}} = 0, S = 1) \\ S_C^{(0)}(x; z, w) &:= \hat{P}(C > x \mid Z = z, W = w, R^{\mathrm{obs}} = 0, S = 1) \end{aligned}$$

33

for $x \in \{x_1,\ldots,x_{m_1}\}$ (because here $(x,z,w) \in \Omega_0$[2]), and

$$S_T^{(1)}(x;z,w) := \hat{P}(T > x \mid Z = z, W = w, R^{\text{obs}} = 1)$$
$$S_C^{(1)}(x;z,w) := \hat{P}(C > x \mid Z = z, W = w, R^{\text{obs}} = 1)$$

for $x \in \{x_{m+1},\ldots,x_n\}$ (because here $(x,z,w) \in \Omega_1$). The survival probabilities (cumulative) are then converted to discrete probability mass as follows. For $(x,z,w) \in \Omega_0$, suppose $x = x_i$ for some $1 \leq i \leq m_1$, then

$$p_T^{(0)}(x;z,w) := \hat{P}(T = x \mid Z = z, W = w, R^{\text{obs}} = 0, S = 1)$$
$$= \begin{cases} 1 - S_T^{(0)}(x_1;z,w) & \text{if } i = 1, \\ S_T^{(0)}(x_{i-1};z,w) - S_T^{(0)}(x_i;z,w) & \text{otherwise.} \end{cases}$$

Similarly, we calculate

$$p_C^{(0)}(x;z,w) := \hat{P}(C = x \mid Z = z, W = w, R^{\text{obs}} = 0, S = 1).$$

The probability on (X,Δ) for Ω_0 is calculated as

$$\hat{P}(X = x, \Delta = \delta \mid R^{\text{obs}} = 0, Z = z, W = w, S = 1)$$
$$= \{p_T^{(0)}(x;z,w) S_C^{(0)}(x;z,w)\}^{\delta} \ \{S_T^{(0)}(x;z,w) p_C^{(0)}(x;z,w)\}^{(1-\delta)}, \qquad \text{(D.2)}$$

and is further normalized over $x \in \{x_1,\ldots,x_{m_1}\}$ (within each (z,w) pair) to ensure

$$\sum_{i=1}^{m_1} \hat{P}(X = x_i, \Delta = \delta_i \mid R^{\text{obs}} = 0, Z = z, W = w, S = 1) = 1. \qquad \text{(D.3)}$$

$p_T^{(1)}(x;z,w)$, $p_C^{(1)}(x;z,w)$, and the probability on (X,Δ) for Ω_1, $\hat{P}(X = x, \Delta = \delta \mid R^{\text{obs}} = 1, Z = z, W = w)$, is calculated similarly.

– When log-normal regression is used as the working model, we fit the four log-normal regressions (D.1) using survreg function in R package survival (Therneau, 2015), without sharing parameters across models. This gives the following prob-

---

[2]We are abusing the notation slightly; more precisely, we meant $(r^{\text{obs}} = 0, z, w, s = 1, x, \delta) \in \Omega_0$.

ability density functions (with respect to Lebesgue measure):

$$f_T^{(0)}(t;z,w) := p_{T|Z,W,R^{\mathrm{obs}}=0,S=1}(t \mid z,w),$$
$$f_C^{(0)}(t;z,w) := p_{C|Z,W,R^{\mathrm{obs}}=0,S=1}(t \mid z,w),$$
$$f_T^{(1)}(t;z,w) := p_{T|Z,W,R^{\mathrm{obs}}=1}(t \mid z,w),$$
$$f_C^{(1)}(t;z,w) := p_{C|Z,W,R^{\mathrm{obs}}=1}(t \mid z,w).$$

These probability density functions are then converted to probability mass functions with positive probability only at points in Ω_0 and Ω_1, respectively. For example, for $(x,z,w) \in \Omega_0$, suppose $x = x_i$ for some $1 \leq i \leq m_1$, then

$$
\begin{aligned}
p_T^{(0)}(x;z,w) &:= \hat{P}(T = x \mid Z = z, W = w, R^{\mathrm{obs}} = 0, S = 1) \\
&= \begin{cases}
\int_0^{(x_1+x_2)/2} f_T^{(0)}(t;z,w)\,dt & \text{if } i = 1, \\
\int_{(x_{i-1}+x_i)/2}^{(x_i+x_{i+1})/2} f_T^{(0)}(t;z,w)\,dt & \text{if } 2 \leq i \leq m_1 - 1, \\
\int_{(x_{m_1-1}+x_{m_1})/2}^{\infty} f_T^{(0)}(t;z,w)\,dt & \text{if } i = m_1.
\end{cases}
\end{aligned}
$$

$p_C^{(0)}(x;z,w)$, $p_T^{(1)}(x;z,w)$ and $p_T^{(1)}(x;z,w)$ are computed similarly. The survival probabilities (such as $S_T^{(0)}(x;z,w)$) are then calculated from the probability mass functions, and the probability on (X,Δ) is calculated as (D.2) and further normalized to ensure (D.3).

E   Additional results regarding the impact of γ on the estimated mortality rate

Here we include results of the estimated mortality rate based on pseudo data sets with various γ values, using three estimators (DE.LN in Figure E.1, KM.S in Figure E.2, KM.C in Figure E.3). The result of the estimator DE.Cox has been presented and discussed in the "*Impact of the double-sampling selection criteria on the estimated mortality rate*" paragraph of Section 5.
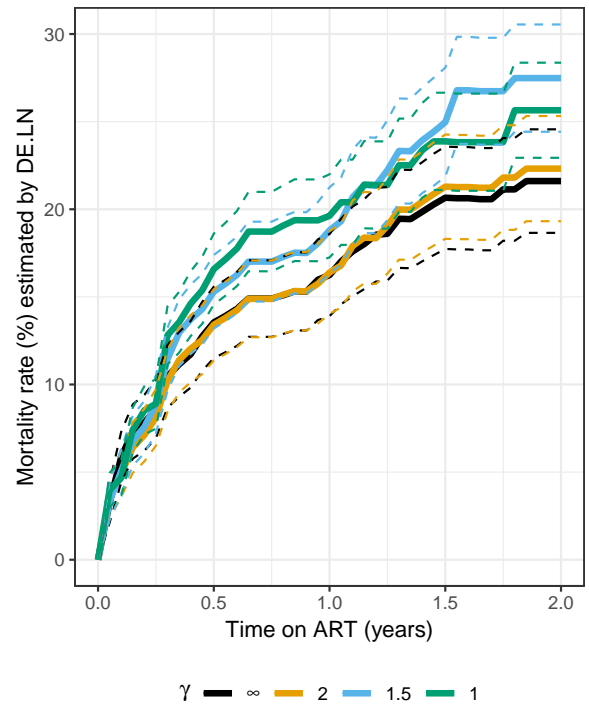
Figure E.1: Estimated mortality rate (solid line) with 95% pointwise confidence interval (dashed line) estimated by DE.LN, using PEPFAR patients data with various γ restrictions: all double-samples included ($\gamma = \infty$), past 2-year dropouts included ($\gamma = 2$), past 1.5-year dropouts included ($\gamma = 1.5$), past 1-year dropouts included ($\gamma = 1$).
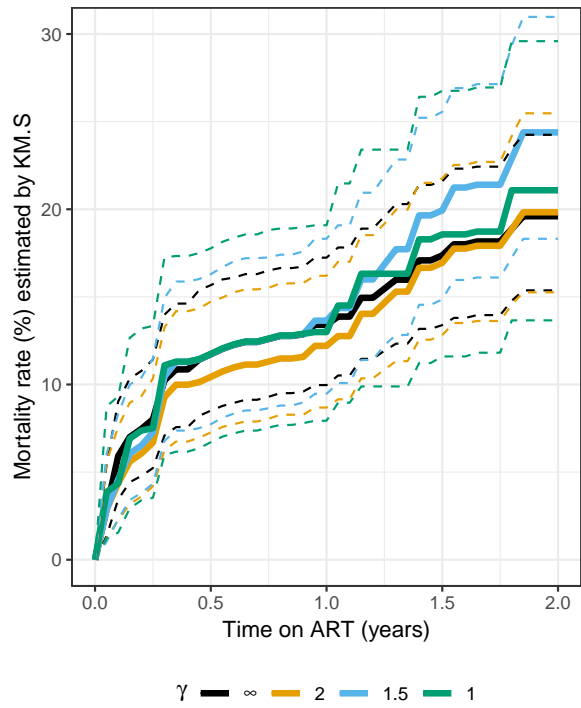
Figure E.2: Estimated mortality rate (solid line) with 95% pointwise confidence interval (dashed line) estimated by KM.S, using PEPFAR patients data with various γ restrictions: all double-samples included ($\gamma = \infty$), past 2-year dropouts included ($\gamma = 2$), past 1.5-year dropouts included ($\gamma = 1.5$), past 1-year dropouts included ($\gamma = 1$).
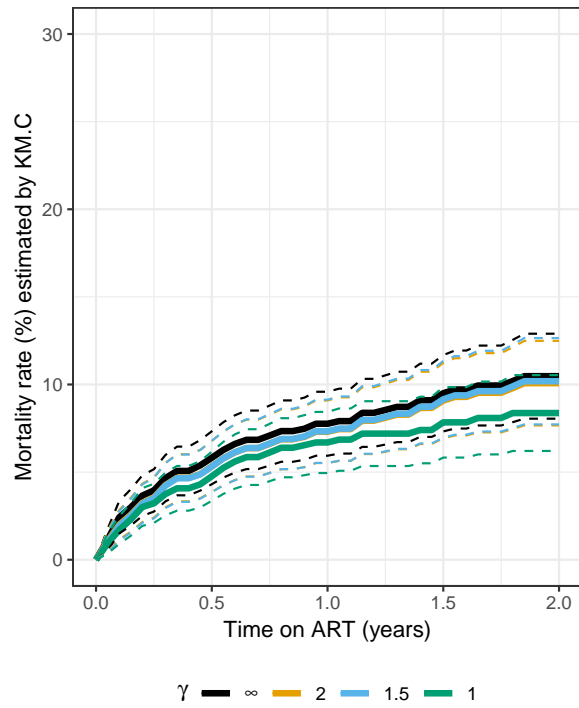
Figure E.3: Estimated mortality rate (solid line) with 95% pointwise confidence interval (dashed line) estimated by KM.C, using PEPFAR patients data with various γ restrictions: all double-samples included ($\gamma = \infty$), past 2-year dropouts included ($\gamma = 2$), past 1.5-year dropouts included ($\gamma = 1.5$), past 1-year dropouts included ($\gamma = 1$).