Journal of Statistical Planning and Inference 216 (2022) 1–14

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Semiparametric inference for merged data from multiple data sources

Takumi Saegusa

Department of Mathematics, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 13 October 2019
Received in revised form 28 April 2021
Accepted 4 May 2021
Available online 12 May 2021

Keywords:
Data integration
Bootstrap
Numerical differentiation
Semiparametric model
Variance estimation

ABSTRACT

We consider general semiparametric inference when data are merged from multiple overlapping sources. Merged data exhibit several characteristics including heterogeneity across multiple data sources, potential unidentified duplicated records, and dependence due to sampling without replacement within each data source. In this paper, we establish a large sample theory for the weighted semiparametric M-estimation with data integration. Our estimator is computable without identifying duplication but corrects bias due to overlapping data sources. The main challenge is that asymptotic variance is not of closed form or contains expectations of unknown functions in general. We propose a novel computational procedure for variance estimation and show its consistency. The finite sample performance is evaluated through a simulation study. A Wilms tumor example is provided.

1. Introduction

Various data sets have become rapidly available thanks to the proliferation of information technology. These data sets, when combined, provide unprecedented opportunities to enhance the quality of inference and accelerate scientific discovery. For instance, reliable analysis on rare scientific phenomena (e.g. rare genetic disease) can be achieved by increasing sample size through incorporating small data sets each of which contains a little information on rare phenomena. If public health data are collected from heterogeneous sources (e.g. clinical trials, disease registries, insurance claims), merging data will significantly reduce selection bias and ensure the generalizability of scientific findings to the broader population. Despite these potential benefits, statistical methodology for data integration has not yet been fully developed in many areas of statistical research.

In this paper, we study semiparametric inference for merged data from multiple sources. A semiparametric model $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ is a collection of probability measures $P_{\theta,\eta}$ dominated by some measure μ and indexed by a finite dimensional parameter $\theta \in \Theta \subset \mathbb{R}^p$ and an infinite-dimensional parameter $\eta \in H \subset (\mathcal{B}, \|\cdot\|)$ where $(\mathcal{B}, \|\cdot\|)$ is a Banach space. When data are independent and identically distributed (i.i.d.), various semiparametric models have been studied including censored regression models (Huang, 1996; Murphy, 1995; Murphy et al., 1997; Parner, 1998), the missing data models (Nan et al., 2009), and the measurement error model (Murphy and van der Vaart, 1996) to name a few (see Bickel et al., 1998; Kosorok, 2008 for more applications). Large sample theory for these models heavily counts on the i.i.d. @assumption but merged data treated in this paper is characterized by (1) heterogeneous data sources, (2) unidentified duplicated records across data sets and (3) dependence due to sampling without replacement. The purpose of this paper is to provide a general inferential procedure to give a basis for studying important semiparametric models in the context of data integration.

E-mail address: tsaegusa@umd.edu.

https://doi.org/10.1016/j.jspi.2021.05.002
0378-3758/

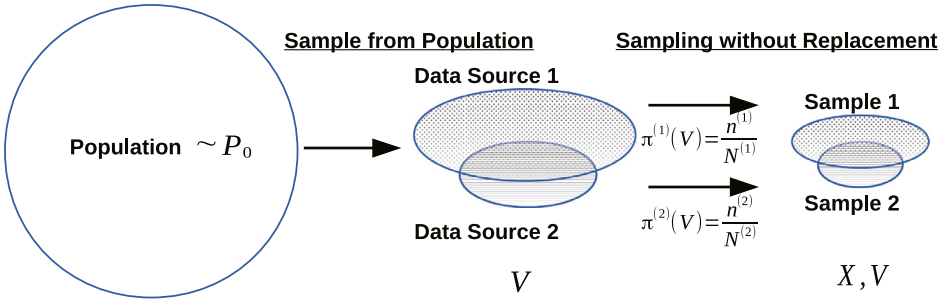

Fig. 1. Sampling scheme for merged data from multiple sources with $J = 2$.

The basic setting considered in this paper is as follows (see also Fig. 1).

• The variables of interest for data integration are a random vector X taking values in a probability space $(\mathcal{X}, \mathcal{A}, P_0)$. The probability measure $P_0 = P_{\theta_0,\eta_0}$ belongs to a statistical model $\mathcal{P}$ with a true parameter (θ_0, η_0).

• Let $V = (\tilde{X}, U) \in \mathcal{V}$ where $\tilde{X}$ is a coarsening of X and U is a vector of auxiliary variables. The variables U do not involve the model $\mathcal{P}$ but help to create data sources. The space $\mathcal{V}$ is composed of J overlapping population data sources $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(J)}$ with $\mathcal{V} = \cup_j \mathcal{V}^{(j)}$ and $\mathcal{V}^{(j)} \cap \mathcal{V}^{(k)} \neq \emptyset$ for some (j, k). Values of V determine membership of data sources.

• Data collection is carried out in a two-stage framework. First, a large i.i.d. sample of $V_1, \ldots, V_N$ is collected from a population. The unit i is assigned to data source j if $V_i \in \mathcal{V}^{(j)}$. Because data sources overlap, the unit i may belong to multiple sources. The sample size of data source $\mathcal{V}^{(j)}$ is denoted as $N^{(j)}$.

• Next, a random sample of size $n^{(j)}$ is selected without replacement from data source $\mathcal{V}^{(j)}$. The selection probability for this data source is $\pi^{(j)}(V_i) = (n^{(j)}/N^{(j)})I\{V_i \in \mathcal{V}^{(j)}\}$ where I is the indicator function. For selected items, variables X are observed.

• The procedure described above is repeated for all data sources. Data sets from each data source are combined and statistical analysis is conducted. If the unit i is selected multiple times, its duplication is not identified.

This two-stage formulation for data collection serves for describing duplicated records in multiple data sets. Duplication naturally occurs in many applications such as public health data research. Clinical and epidemiological studies identify target populations by the inclusion and exclusion criteria. When national disease registries are combined with these studies, diseased subjects must be in a national database as well. Duplicated records are difficult to identify in practice because key identifiers such as names and addresses are often not disclosed for privacy protection in public health data.

Examples covered by our framework include opinion polls (Brick et al., 2006), public health surveillance (Hu et al., 2011), health interview surveys (Cervantes et al., 2006), and the synthesis of existing clinical and epidemiological studies with surveys, disease registries, and healthcare databases (Chatterjee et al., 2016; Keiding and Louis, 2016; Metcalf and Scott, 2009). For the ease of better understanding of our mathematical setting, consider combining a cohort study and a disease registry as a hypothetical example. Variables of interest are disease status D, age A, and a biomarker Z which form $X = (D, A, Z)$. The statistical model $\mathcal{P}$ of interest is the logistic regression model with outcome D and covariates A and Z. The cohort study targets a high risk group defined by age older than 40 and a positive test result $\tilde{Z} \in \{0, 1\}$ of an error-prone inexpensive medical test to measure Z in the state in the United States. Let $S \in \{0, 1\}$ be an indicator of living in the target state. The national disease registry collects information on diseased subjects in the United States. Because disease status, the age category, and potentially mismeasured biomarker are partial information needed for logistic regression, we can write $\tilde{X} = (D, I\{A \geq 40\}, \tilde{Z})$. Since a living address is not considered as a risk factor, the auxiliary variable is $U = S$. Based on $V = (\tilde{X}, U)$, data sources for the cohort study and the disease registry are $\mathcal{V}^{(1)} = \{V : A \geq 40, \tilde{Z} = 1, S = 1\}$ and $\mathcal{V}^{(2)} = \{V : D = 1\}$, respectively. The cohort study collects $n^{(1)}$ by sampling without replacement and then ascertain the disease status D and measure biomarker Z with a more precise medical test. At the diagnosis of disease, information on risk factors such as a biomarker is sent to the disease registry for all diseased subjects.

Merged data is considered to be a biased and dependent sample with duplication. Bias in merged data arises in two ways. Certain data sources are over/under-represented due to biased sampling with different selection probabilities $\pi^{(j)}$ which yields heterogeneity in integrated data. Duplicated records enter the final merged data without identification. There are two types of dependence in merged data. Dependence between data sources is induced through duplicated records from overlapping data sources. Dependence within data sources is generated from sampling without replacement from each data source. These characteristics grossly differentiate our data integration problem from the analysis of an i.i.d. sample. Estimation and the corresponding asymptotic theory require different theory and methods that specialize in data integration in order to address challenging issues of bias, dependence, and duplication.

Data integration problems described above were first studied by our previous work (Saegusa, 2019). We developed a weighting procedure to study the infinite-dimensional M-estimator. The estimator is computable without identifying duplicated records but corrects bias due to duplication and biased sampling. For its asymptotic theory, several specialized probabilistic tools were developed including the uniform law of large numbers and uniform central limit theorem for data integration. Semiparametric estimation was briefly mentioned as an example (Example 5.2 of Saegusa (2019)) but

its asymptotic properties were presented without proof. For asymptotic variance of the M-estimator, a plug-in estimator was proposed and examined in simulation studies.

In this paper, we study weighted semiparametric likelihood estimators proposed by our previous work (Saegusa, 2019). We provide a rigorous large sample theory to establish asymptotic distributions of our semiparametric estimators. The main contribution of this paper is a novel computational procedure to estimate asymptotic variance. The previously proposed plug-in variance estimator has limited uses because asymptotic variance in many semiparametric models do not have a closed form or contain expectations of unknown functions even in the i.i.d. setting (Murphy and van der Vaart, 1999). A popular alternative is resampling methods such as bootstrap (Efron, 1979) and jackknife (Quenouille, 1949; Tukey, 1958). Various kinds of resampling methods cover different data generating mechanisms (see e.g. Shao and Tu, 1995) but existing methods do not address heterogeneity of data sets and duplicated selection. Another approach in the i.i.d. setting is to estimate an efficient information matrix (Murphy and van der Vaart, 1999; Zhang et al., 2010). This approach focuses on an efficient estimator whose asymptotic variance is the inverse of the efficient information. In our setting, the loss in efficiency is expected compared to the i.i.d. setting, and hence estimation of the efficient information does not lead to consistent variance estimation.

Applications of semiparametric inference to data integration problems are largely hampered by the lack of valid variance estimators. Our proposed method is the first to address the challenging issue of estimating complicated asymptotic variance due to heterogeneity and duplicated selection arising from data integration problems. To address the lack of a simple asymptotic variance formula, we adopt two computational methods to estimate different parts of variance. The proposed methodology covers many semiparametric models and is easy to apply without computing complicated asymptotic variance.

The rest of the paper is organized as follows. In Section 2, we introduce our estimator and present its asymptotic properties. Section 3 concerns a novel variance estimation method. Consistency of the proposed estimator is provided. The finite sample performance of our estimator is presented in Section 4. Data example from the national Wilms tumor study is discussed in Section 4. All proofs for theorems in Section 3 and auxiliary results are collected in Section 6. Section 7 concludes with discussion on the future research.

2. Sampling and estimator

We introduce additional notation. Let $R_i^{(j)} \in \{0, 1\}$ be the selection indicator from data source $\mathcal{V}^{(j)}$ for the item i. The selection indicators for the item i is $R_i = (R_i^{(1)}, \ldots, R_i^{(J)})$ where $R_i^{(j)} = 0$ if the item i does not belongs to source $\mathcal{V}^{(j)}$. Selection indicators $R_i^{(j)}$s with items i in data source $\mathcal{V}^{(j)}$ follow the distribution of sampling without replacement where $n^{(j)}$ items are selected out of $N^{(j)}$ items. Since data collection are carried out independently, selection indicators $(R_1^{(j)}, \ldots, R_N^{(j)})$ and $(R_1^{(k)}, \ldots, R_N^{(k)})$ for different selection processes are conditionally independent given $V_1, \ldots, V_N$ if $j \neq k$. For $V \in \mathcal{V}^{(j)}$, we assume the selection probability $\pi^{(j)}(V) = n^{(j)}/N^{(j)}$ converges to $p_j > 0$ as $N \to \infty$. Since $n^{(j)}$ is at the disposal of a study investigator for the jth data source, this convergence is deterministic. We write the membership probability in source $\mathcal{V}^{(j)}$ as $\nu^{(j)} = P(V \in \mathcal{V}^{(j)})$ and the conditional expectation and conditional variance given membership in source $\mathcal{V}^{(j)}$ as $E^{(j)}$ and $\mathrm{Var}^{(j)}$.

The important assumption we make is that additional membership in data sources can be identified for selected items. In other words, if the item is selected from some data source, we assume that we can obtain information on the other data sources the item belongs to. We do not assume the knowledge of data source membership for non-selected items nor knowledge of actual selections for items selected from at least one data source. This assumption is not too restrictive because a target population of each data collection process is clearly specified in general. For public health data integration, one can compare different inclusion and exclusion criteria to identify additional data source membership without seeking information on multiple selection. If necessary, one can ask additional question on characteristics of the selected items. For telephone surveys using landline and cell-phone lists, one can always ask a cell phone user if she has a landline phone as well in order to identify her membership in data source of landline users.

The desirable properties that our estimator holds are that (1) correction of bias due to biased sampling and duplication, and that (2) computability without identification of duplicated records. To describe our estimator, we begin with $J = 2$ data sources. The important component of our estimator is

$$\rho(v) = (\rho^{(1)}(v), \rho^{(2)}(v)) \equiv \begin{cases} (1, 0) & \text{if} \quad v \in \mathcal{V}^{(1)} \text{ and } v \notin \mathcal{V}^{(2)}, \\ (0, 1) & \text{if} \quad v \notin \mathcal{V}^{(1)} \text{ and } v \in \mathcal{V}^{(2)}, \\ (c^{(1)}, c^{(2)}) & \text{if} \quad v \in \mathcal{V}^{(1)} \cap \mathcal{V}^{(2)}, \end{cases}$$

for positive constants $c^{(1)}, c^{(2)}$ with $c^{(1)} + c^{(2)} = 1$. The evaluation of this function only requires the membership in the mutually exclusive subsets of $\mathcal{V}$ based on data sources $\mathcal{V}^{(1)}$ and $\mathcal{V}^{(2)}$. We can compute the value of ρ for selected items because information on data source membership is available for selected items. The choice of ρ is at the disposal of a data analyst but different choice yields different asymptotic variance. We follow the optimal choice of ρ considered by Proposition 3.1 of Saegusa (2019) in simulation and data analysis below. Let $p_{\theta,\eta} = dP_{\theta,\eta}/d\mu$ be a density of $P_{\theta,\eta}$ and $\ell_{\theta,\eta} = \log p_{\theta,\eta}$. The weighted log likelihood is then computed by

$$\mathbb{P}_N^H \ell_{\theta,\eta} \equiv \frac{1}{N} \sum_{i=1}^{N} \left(\frac{R_i^{(1)} \rho^{(1)}(V_i)}{\pi^{(1)}(V_i)} + \frac{R_i^{(2)} \rho^{(2)}(V_i)}{\pi^{(2)}(V_i)} \right) \ell_{\theta,\eta}(X_i).$$

Here we use the convention $0/0 = 0$ for $R^{(j)}/\pi^{(j)}(V)$. This weighted average of the log likelihood is unbiased for the expectation of the log density because the inverse probability weights $R^{(j)}/\pi^{(j)}(V)$ have expectation 1 and $\rho^{(1)}(v)+\rho^{(2)}(v) = 1$ for any v by definition. Moreover, it is computable without identifying duplication because two terms in $\mathbb{P}_N^H \ell_{\theta,\eta}$ denoted as

$$\mathbb{P}_N^H \ell_{\theta,\eta} = \frac{1}{N}\sum_{i=1}^N \frac{R_i^{(1)}\rho^{(1)}(V_i)}{\pi^{(1)}(V_i)}\ell_{\theta,\eta}(X_i) + \frac{1}{N}\sum_{i=1}^N \frac{R_i^{(2)}\rho^{(2)}(V_i)}{\pi^{(2)}(V_i)}\ell_{\theta,\eta}(X_i)$$

can be computed separately for different data collection processes.

The extension to more than two data sources is easily obtained by introducing $\rho(v) = (\rho^{(1)}(v), \ldots, \rho^{(J)}(v))$ given by

$$\rho^{(j)}(v) = \begin{cases} 1, & v \in \mathcal{V}^{(j)} \cap \left(\cup_{m\neq j}\mathcal{V}^{(m)}\right)^c, \\ c^{(j)}_{k_1,\ldots,k_l}, & v \in \mathcal{V}^{(j)} \cap \left(\cap_{m=1}^l \mathcal{V}^{(k_m)}\right) \cap \left(\cup_{m\notin\{j,k_1,\ldots,k_l\}}\mathcal{V}^{(m)}\right)^c, \\ 0, & v \notin \mathcal{V}^{(j)}, \end{cases}$$

with $j, k_1, \ldots, k_l$ all different and $\sum_{j=1}^J \rho^{(j)}(v) = 1$. The weighted likelihood is

$$\mathbb{P}_N^H \ell_{\theta,\eta} = \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^J \frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)}\ell_{\theta,\eta}(X_i).$$

The weighting by both inverse probability weights and ρ was originally proposed by Hartley (1962) and Hartley (1974) in survey sampling and then applied to data integration problems by Saegusa (2019). Throughout we denote the weighted average of $f(X)$ by $\mathbb{P}_N^H f$ for different functions f as follows. Let

$$W_{Ni} \equiv \sum_{j=1}^J \frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)}$$

be the weight for the ith item. For the function f defined on $\mathcal{X}$, the weighted average $\mathbb{P}_N^H f$ is formally defined as

$$\mathbb{P}_N^H f \equiv \frac{1}{N}\sum_{i=1}^N W_{Ni} f(X_i) = \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^J \frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)} f(X_i).$$

Here we follow the convention that the variable X is suppressed in $\mathbb{P}_N^H f$. We write the weighted empirical process by $\mathbb{G}_N^H = \sqrt{N}(\mathbb{P}_N^H - P_0)$. This stochastic process evaluated at f is $\mathbb{G}_N^H f = \sqrt{N}(\mathbb{P}_N^H f - P_0 f)$ where $P_0 f$ is the expectation of $f(X)$ with respect to P_0.

The proposed estimator is the maximizer of the weighted likelihood:

$$(\hat\theta_N, \hat\eta_N) = \underset{\theta\in\Theta,\eta\in H}{\arg\max}\, \mathbb{P}_N^H \ell_{\theta,\eta}. \tag{1}$$

In the i.i.d. setting, the maximum likelihood estimators (MLE) obtained as the maximizer of the likelihood often solves the semiparametric likelihood equations, exactly or approximately. Thus, we assume the proposed estimator solves the weighted semiparametric likelihood equations given by

$$\begin{aligned} &\Psi_{N,1}(\theta,\eta) \equiv \mathbb{P}_N^H \dot\ell_{\theta,\eta} = o_P(N^{-1/2}), \\ &\Psi_{N,2}(\theta,\eta) \equiv \mathbb{P}_N^H (B_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h) = o_P(N^{-1/2}), \end{aligned} \tag{2}$$

for every $h \in \mathcal{H}$. This assumption can be verified as in the i.i.d. setting (see e.g. Huang, 1996). Here $\dot\ell_{\theta,\eta}$ is the score function for θ obtained as the derivative of $\log p_{\theta,\eta}$ with respect to θ and the score operator $B_{\theta,\eta}$ is the bounded linear operator mapping a direction h in some Hilbert space $\mathcal{H}$ of one-dimensional submodels along which η passes through η_0 to a square integrable random variable with respect to $P_{\theta,\eta}$. The score operator is motivated by the observation that a semiparametric model is an infinitely many collection of one-dimensional parametric submodels such that score can be computed for each submodel that corresponds to h. We denote the adjoint operator of $B_{\theta,\eta}$ as $B^*_{\theta,\eta}$. For more details, see van der Vaart (1998).

We assume the following regularity conditions.

Condition 1 (*Asymptotic Equicontinuity*)**.** *Let*

$$\begin{aligned} &\mathcal{F}_1(\delta) = \{\dot\ell_{\theta,\eta} : |\theta-\theta_0| + \|\eta-\eta_0\| < \delta\}, \\ &\mathcal{F}_2(\delta) = \{B_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h : \ h\in\mathcal{H}, |\theta-\theta_0| + \|\eta-\eta_0\| < \delta\}. \end{aligned}$$

There exists a $\delta_0 > 0$ such that (a) $\mathcal{F}_k(\delta_0)$, $k = 1, 2$, are P_0-Donsker classes and $\sup_{h\in\mathcal{H}} P_0|f_j - f_{0,j}|^2 \to 0$, as $|\theta-\theta_0| + \|\eta-\eta_0\| \to 0$, for every $f_j \in \mathcal{F}_j(\delta_0)$, $j = 1, 2$, where $f_{0,1} = \dot\ell_{\theta_0,\eta_0}$ and $f_{0,2} = B_{\theta_0,\eta_0}h - P_0 B_{\theta_0,\eta_0}h$, and that (b) $\mathcal{F}_k(\delta_0)$, $k = 1, 2$, have integrable envelopes.

Condition 2 (*Partition of the Estimating Function*). *The map* $\Psi = (\Psi_1, \Psi_2) : \Theta \times H \mapsto \mathbb{R}^p \times \ell^\infty(\mathcal{H})$ *with components*

$$\Psi_1(\theta, \eta) \equiv P_0 \Psi_{N,1}(\theta, \eta) = P_0 \dot{\ell}_{\theta,\eta},$$
$$\Psi_2(\theta, \eta)h \equiv P_0 \Psi_{N,2}(\theta, \eta)h = P_0 B_{\theta,\eta} h - P_{\theta,\eta} B_{\theta,\eta} h,$$

with $h \in \mathcal{H}$ *has a continuously invertible Fréchet derivative map* $\dot{\Psi}_0 = (\dot{\Psi}_{11}, \dot{\Psi}_{12}, \dot{\Psi}_{21}, \dot{\Psi}_{22})$ *at* (θ_0, η_0) *given by* $\dot{\Psi}_{ij}(\theta_0, \eta_0)h = P_0(\dot{\psi}_{i,j,\theta_0,\eta_0,h})$, $i, j \in \{1, 2\}$ *in terms of* $L_2(P_0)$*-derivatives of* $\psi_{1,\theta,\eta,h} = \dot{\ell}_{\theta,\eta}$ *and* $\psi_{2,\theta,\eta,h} = B_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h$*; that is,*

$$\sup_{h\in\mathcal{H}} [P_0\{\psi_{i,\theta,\eta_0,h} - \psi_{i,\theta_0,\eta_0,h} - \dot{\psi}_{i1,\theta_0,\eta_0,h}(\theta - \theta_0)\}^2]^{1/2} = o(\|\theta - \theta_0\|),$$
$$\sup_{h\in\mathcal{H}} [P_0\{\psi_{i,\theta_0,\eta,h} - \psi_{i,\theta_0,\eta_0,h} - \dot{\psi}_{i2,\theta_0,\eta_0,h}(\eta - \eta_0)\}^2]^{1/2} = o(\|\eta - \eta_0\|).$$

Furthermore, $\dot{\Psi}_0$ *admits a partition*

$$(\theta - \theta_0, \eta - \eta_0) \mapsto \begin{pmatrix} \dot{\Psi}_{11} & \dot{\Psi}_{12} \\ \dot{\Psi}_{21} & \dot{\Psi}_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \eta - \eta_0 \end{pmatrix},$$

where

$$\dot{\Psi}_{11}(\theta - \theta_0) = -P_{\theta_0,\eta_0} \dot{\ell}_{\theta_0,\eta_0} \dot{\ell}^T_{\theta_0,\eta_0} (\theta - \theta_0),$$
$$\dot{\Psi}_{12}(\eta - \eta_0) = -\int B^*_{\theta_0,\eta_0} \dot{\ell}_{\theta_0,\eta_0} d(\eta - \eta_0),$$
$$\dot{\Psi}_{21}(\theta - \theta_0)h = -P_{\theta_0,\eta_0} B_{\theta_0,\eta_0} h \dot{\ell}^T_{\theta_0,\eta_0} (\theta - \theta_0),$$
$$\dot{\Psi}_{22}(\eta - \eta_0)h = -\int B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0} h d(\eta - \eta_0),$$

and $B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0}$ *is continuously invertible.*

The efficient score for θ *in the i.i.d. setting is denoted as*

$$\ell^*_0 = (I - B_{\theta_0,\eta_0}(B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0})^{-1} B^*_{\theta_0,\eta_0}) \dot{\ell}_{\theta_0,\eta_0}.$$

The corresponding efficient information and efficient influence function for θ *are denoted as* $\tilde{I}_0 = P_0[\ell^*_0(\ell^*_0)^T]$ *and* $\tilde{\ell}_0 = \tilde{I}_0^{-1}\ell^*_0$ *respectively.*

Now we present the result of asymptotic normality of the proposed estimator. The following theorem is an extension of the semiparametric Z-theorem of van der Vaart (1998) (Theorem 25.90, pages 420–421) in the i.i.d. setting to data integration.

Theorem 2.1. *Suppose* $(\hat{\theta}_N, \hat{\eta}_N)$ *solves semiparametric likelihood equations* (2) *and is consistent for* (θ_0, η_0)*. Under Conditions* 1 *and* 2*,*

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbb{G}_N^H \tilde{\ell}_0 + o_P(1) \rightsquigarrow_d Z \sim N(0, \Sigma)$$

where

$$\Sigma \equiv \tilde{I}_0^{-1} + \sum_{j=1}^{J} \nu_j \frac{1 - p_j}{p_j} \mathrm{Var}^{(j)}(\rho^{(j)} \tilde{\ell}_0). \tag{3}$$

If we would obtain the i.i.d. sample of size N, the MLE achieves the information bound $\tilde{I}_0^{-1}$. Additional sampling from each data source results in the increase of asymptotic variance for our weighted likelihood estimator. If all items are selected from data sources (i.e., the limit of selection probabilities $p^{(j)} = 1, j = 1, \ldots, J$), then asymptotic variance is the same as $\tilde{I}_0^{-1}$. Large data sources (i.e., large $\nu^{(j)}$) contribute more to asymptotic variance. Recall that information on duplication is contained in the variable $\rho(V) = (\rho^{(1)}(V), \ldots, \rho^{(J)}(V))$ in the proposed estimation procedure. The effect of duplication appears through $\rho^{(j)}$ in the conditional variance.

3. Variance estimation

In Section 2, we derived the asymptotic variance Σ of the proposed estimator $(\hat{\theta}_N, \hat{\eta}_N)$ in (3). As discussed before, the efficient information bound $\tilde{I}_0^{-1} = \mathrm{Var}(\tilde{\ell}_0)$ may not be of closed form or contains the expectation of unknown functions in many semiparametric models so that plug-in variance estimators are not available in the i.i.d. setting. This issue remains in our data integration setting both through $\tilde{I}_0^{-1} =$ and through $\mathrm{Var}_0^{(j)}\{\rho^{(j)}(V)\tilde{\ell}_0(X)\}$ in the asymptotic variance Σ. To address this challenging issue, we adopt two computational methods to separately estimate different parts of Σ.

To motivate our proposed variance estimation methodology, we decompose the asymptotic variance $\Sigma = \Sigma_1 + \Sigma_2$ into *population variance* Σ_1 and *design variance* Σ_2 given by

$$\Sigma_1 = \tilde{I}_0^{-1}, \quad \Sigma_2 = \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \mathrm{Var}_0^{(j)}\{\rho^{(j)}(V)\tilde{\ell}_0(X)\}.$$

Recall the two-stage formulation of data integration. If we have hypothetically access to the complete information on X instead of V at the first stage of sampling from the infinite population, the analysis of the i.i.d. sample $X_1, \ldots, X_N$ with maximum likelihood estimation yields asymptotic variance Σ_1 only. Because we only observe V at the first stage, sampling from the finite population $V_1, \ldots, V_N$ from data sources and merging overlapping data sets yield the additional variance at the second stage. This observation suggests the following approach. For the estimation of Σ_1, some existing method (or its modification) developed for the i.i.d. sample is potentially useful. Because randomness in X is captured by Σ_1, a promising method to estimate Σ_2 is a method developed in survey sampling that targets randomness in R arising from selections from the finite population.

3.1. Efficient information

Estimating population variance $\Sigma_1 = \tilde{I}_0^{-1}$ reduces to the estimation of the efficient information matrix $\tilde{I}_0$. As the Fisher information matrix is computed as the expectation of the score multiplied by its transpose, the efficient information is obtained as the expectation of the efficient score multiplied by its transpose. That is, $\tilde{I}_0 = P_0[\ell_0^*(\ell_0^*)^T]$. As seen from the form of the efficient score ℓ_0^* in Condition 2, a derivative $\dot{\ell}_{\hat{\theta}_N, \hat{\eta}_N}$ of the log likelihood with respect to θ at $(\hat{\theta}_N, \hat{\eta}_N)$ does not estimate ℓ_0^*. Estimating ℓ_0^* directly by a plug-in estimator is also difficult in general since $B_{\theta,\eta}^*$ in ℓ_0^* involves the expectation with respect to an unknown probability measure.

Our approach is to directly estimate $\tilde{I}_0$ without estimating the efficient score ℓ_0^*. Recall that a semiparametric model is the infinitely many collection of finite dimensional parametric submodels for which the Fisher information can be computed. The efficient information is defined as the infimum of the Fisher information over all finite dimensional submodels. The efficient score thus corresponds to the score for an unknown finite dimensional submodel with the smallest Fisher information, or the least favorable submodel (Murphy and van der Vaart, 1999, 2000) under regularity conditions. If we knew the least favorable submodel, the efficient information can be estimated from the negative average of the second derivative of the corresponding log likelihood. The least favorable submodel is, unfortunately, unknown and difficult to estimate. Instead, Murphy and van der Vaart (1999) found the likelihood based on the least favorable submodel can be approximated by the profile likelihood that profiles out the infinite-dimensional nuisance parameter. Because we do not know how the least favorable submodel is parametrized, we cannot take a derivative of the profile likelihood with respect to unknown parameters in the submodel. Instead, Murphy and van der Vaart (1999) proposed a numerical second derivative of the profile likelihood to estimate the efficient information in the i.i.d. setting. We extend this methodology to merged data from overlapping sources.

Let $\mathbb{M}_N(\theta)$ be the weighted profile likelihood

$$\mathbb{M}_N(\theta) \equiv \sup_{\eta \in H} \mathbb{P}_N^H \ell_{\theta,\eta},$$

obtained by maximizing the weighted likelihood with respect to the infinite-dimensional parameter η for a fixed θ. The profile likelihood may not have a closed form in general because its computation involves an infinite-dimensional optimization problem for η. Theoretical properties of the semiparametric profile likelihood were established by Murphy and van der Vaart (2000) in the i.i.d. setting but its stochastic behavior has not been studied in our context. In this paper, we focus on the claim that the weighted profile likelihood $\mathbb{M}_N$ well approximates the weighted log likelihood based on the least favorable submodel as in the i.i.d. setting. We assume similar conditions considered by Murphy and van der Vaart (1999). These conditions were verified by Murphy and van der Vaart (1999, 2000) in several semiparametric models in the i.i.d. setting. The same proof techniques continue to work for our setting with the help of empirical process results by Saegusa (2019) for data integration.

Condition 3 (*Least Favorable Submodel*). *(a) Denote $\gamma \equiv (\theta, \eta)$, and, for a fixed θ, $\hat{\gamma}_{\theta,N} \equiv (\theta, \hat{\eta}_{\theta,N})$ with $\hat{\eta}_{\theta,N} \equiv \arg\max_{\eta \in H} \mathbb{P}_N^{\pi} \ell_{\theta,\eta}$. For every $\tilde{\theta}_N \to_P \theta_0$, $\hat{\eta}_{\tilde{\theta}_N,N} \to_P \eta_0$.*

(b) For each $\gamma = (\theta, \eta)$ there exists a map $t \mapsto \boldsymbol{\eta}_t(\gamma)$ from a neighborhood of θ to the parameter set H for η such that (i) the map $t \mapsto \ell(t, \gamma)(x)$ defined by

$$\ell(t, \gamma)(x) = \ell_{t, \boldsymbol{\eta}_t(\gamma)}(x)$$

is twice continuously differentiable for every $x \in \mathcal{X}$ where the first and second derivatives are denoted by $\dot{\ell}(t, \gamma)(x)$, and $\ddot{\ell}(t, \gamma)(x)$, respectively, and such that (ii) a submodel with parameters $(t, \boldsymbol{\eta}_t(\gamma))$ passes through $\gamma = (\theta, \eta)$ at $t = \theta$;

$$\boldsymbol{\eta}_\theta(\theta, \eta) = \eta, \quad \textit{every} \quad (\theta, \eta). \tag{4}$$

(c) For any random sequences $\tilde{\theta}$ *and* $\overline{\gamma}$ *such that* $\tilde{\theta} \to_P \theta_0$ *and* $\overline{\gamma} \to_P \gamma_0$,

$$\mathbb{G}_N^H \dot{\ell}(\tilde{\theta}, \overline{\gamma}) = \mathbb{G}_N^H \ell_0^* + o_P(1), \tag{5}$$

$$\mathbb{P}_N^H \ddot{\ell}(\tilde{\theta}, \overline{\gamma}) \to_P -\tilde{I}_0, \tag{6}$$

$$P_0 \dot{\ell}(\tilde{\theta}, \hat{\gamma}_{\tilde{\theta},N}) = -\tilde{I}_0(\tilde{\theta} - \theta_0) + o_P\left(\|\tilde{\theta} - \theta_0\| + N^{-1/2}\right). \tag{7}$$

Under the above condition, numerical second derivative of the profile likelihood $\mathbb{M}_N$ is consistent for elements of the efficient information matrix $\tilde{I}_0$.

Theorem 3.1. *Suppose that the same conditions in Theorem 2.1 and Condition 3 hold. Then*

$$-2\frac{\mathbb{M}_N(\hat{\theta}_N + h_N v_N) - \mathbb{M}_N(\hat{\theta}_N)}{h_N^2} \to_P v_0^T \tilde{I}_0 v_0, \tag{8}$$

for every random sequence $h_N \to_P 0$ *such that* $(\sqrt{N}h_N)^{-1} = O_P(1)$ *and for every sequence* $v_N \to_P v_0$.

It is straightforward to estimate $\tilde{I}_0$ using the numerical second derivatives of $\mathbb{M}_N$. For example, (i,j)-elements $\tilde{I}_{i,j,0}$ of $\tilde{I}_0$ with $i,j = 1,2$, can be obtained by setting $v_N = (1,0,\ldots,0)$, $v_N = (0,1,0,\ldots,0)$, and $v_N = (1,1,0,0,\ldots,0)$ to estimate $\tilde{I}_{1,1,0}$, $\tilde{I}_{2,2,0}$, and $\tilde{I}_{1,1,0} + 2\tilde{I}_{1,2,0} + \tilde{I}_{2,2,0}$ respectively. In practice one can set $h_N = N^{-1/2}$ so that $\sqrt{N}h_N = 1$. Once we obtain the estimate of $\tilde{I}_0$, we estimate the population variance by computing its inverse.

3.2. Sampling from data sources

For the estimation of design variance Σ_2, we apply Gross' bootstrap (Gross, 1980; Bickel and Freedman, 1984) to each data set selected from the same sampling procedure (i.e., the jth data set consisting of all items i in $\mathcal{V}^{(j)}$ with $R_i^{(j)} = 1$). This bootstrap method was originally proposed for the analysis of stratified samples in survey sampling. Because the main focus of survey sampling is finite population, this method reproduces the randomness in R due to sampling from the finite population where collected variables X are treated as non-random. This feature is suitable for our setting because design variance Σ_2 represents randomness due to sampling from data sources given the finite population $V_1,\ldots,V_N$.

To adjust the bootstrap method for stratified samples to merged data with duplication, we consider applying Gross' bootstrap to sampling from each data source separately. The following procedure mimics sampling from data source j. Because unselected observations with $R^{(j)} = 0$ are generally not available, the bootstrap population must be created from selected observations with $R^{(j)} = 1$. Recall that from the data source $\mathcal{V}^{(j)}$ of size $N^{(j)}$, $n^{(j)}$ items are selected without replacement. Let $k^{(j)} \in \mathbb{N}$ and $r^{(j)} \in \mathbb{N}$ be divisor and remainder such that $N^{(j)} = n^{(j)}k^{(j)} + r^{(j)}$. For items with $R^{(j)} = 1$, their $k^{(j)}$ copies or $k^{(j)} + 1$ copies are created to form the bootstrap population j with probability $s^{(j)}$ and $1 - s^{(j)}$ where $s^{(j)} = (1 - r^{(j)}/n^{(j)})\{1 - r^{(j)}/(N^{(j)} - 1)\}$. Then $n^{(j)}$ items are selected from the bootstrap population without replacement into the bootstrap sample j. As a simple example, suppose we select 5 items out of 10. In this case, $n = 5, N = 10, k = 2, r = 0, s = 1$. We create copies of 5 selected items to create a bootstrap population of size 10. Then we sample 5 items without replacement. If $n = 4$ and $N = 10$, then $r = 2$ and $s = 7/18$. We choose a bootstrap population of size 8 created by 2 copies of selected items with probability $s = 7/18$ or a bootstrap population of size 12 created by 3 copies of selected items with probability $1 - s$. Then we sample 4 items without replacement.

This bootstrap procedure is repeated for all data sets sampled from each data source. Let $B_i^{(j)}$ be the count that the ith items are selected in the bootstrap sample j. Because the bootstrap population contains multiple copies of selected items, $B_i^{(j)}$ can be 0,1, …, or $\max\{k^{(j)}, (k^{(j)} + 1)I\{r^{(j)} > 0\}\}$. The bootstrap weighted likelihood is now defined as

$$\hat{\mathbb{P}}_N^H \ell_{\theta,\eta} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} B_i^{(j)} \frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)} \ell_{\theta,\eta}(X_i).$$

The bootstrap estimator is the maximizer of the bootstrap weighted likelihood:

$$(\hat{\theta}_N^*, \hat{\eta}_N^*) = \underset{\theta\in\Theta,\eta\in H}{\arg\max}\, \hat{\mathbb{P}}_N^H \ell_{\theta,\eta}.$$

This bootstrap estimator is assumed to solve (exactly or approximately) the bootstrap semiparametric likelihood equations given by

$$\hat{\mathbb{P}}_N^H \dot{\ell}_{\theta,\eta} = o_P(N^{-1/2}),$$
$$\sup_{h\in\mathcal{H}} |\hat{\mathbb{P}}_N^H (B_{\theta,\eta} h - P_{\theta,\eta} B_{\theta,\eta} h)| = o_P(N^{-1/2}). \tag{9}$$

For the estimation of Σ_2, we generate B bootstrap samples to compute the B bootstrap weighted likelihood estimators $\hat{\theta}_{N,1}^*,\ldots,\hat{\theta}_{N,B}^*$ of θ as described above. The proposed bootstrap estimator of Σ_2 is their sample variance.

Independent applications of Gross' bootstrap to each data set change the number of duplications from one bootstrap sample to another. This issue may raise a concern about consistent estimation of Σ_2 by our bootstrap procedure, but

our method is based on the careful examination of the uniform central limit theorem for data integration (Theorem 3.2 of Saegusa (2019)). As seen in Theorem 2.1, asymptotic normality of $\hat{\theta}_N$ reduces to the central limit theorem applied to $\mathbb{G}_N^H \tilde{\ell}_0$. When applying Theorem 3.2 of Saegusa (2019), the weighted average $\mathbb{G}_N^H \tilde{\ell}_0$ is decomposed into $J+1$ terms (see the last display of page 8 of Saegusa (2019)). The first term corresponds to Σ_1. The last J terms which correspond to Σ_2 are weighted averages defined on each data source. Although these J weighted averages have duplications, they are shown to be uncorrelated so that central limit theorem applied to each average yields independent normal random variables in the limit (see the proof of Theorem 3.2 of Saegusa (2019)). This observation provides a heuristic justification of independent applications of Gross' bootstrap to each data source. For a rigorous proof, see Section 6. Now, we obtain the following bootstrap Z-theorem.

Theorem 3.2. *Suppose $(\hat{\theta}_N^*, \hat{\eta}_N^*)$ solves bootstrap semiparametric likelihood equations (9) and is consistent for (θ_0, η_0). Under Conditions 1 and 2,*

$$\sqrt{N}(\hat{\theta}_N^* - \hat{\theta}_N) \rightsquigarrow_d \quad Z_2 \sim N(0, \Sigma_2),$$

conditionally on data.

4. Numerical example

We evaluate the finite sample property of the proposed weighted likelihood estimator through simulation studies using the Cox proportional hazards model with right censoring. Let T be a time to event from the Cox model with the hazard function given a vector of covariate $Z = z$

$$\lambda(t|z) = \lambda(t)\exp(z^T\theta)$$

where $\lambda(t)$ is a baseline hazard function and θ is a regression parameter. Under right censoring, the censored failure time $Y = \min\{T, C\}$ and the censoring indicator $\Delta = I\{T \le C\}$ are observed where C is a censoring variable. The conditional independence of T and C given Z is assumed. The weighted log likelihood is given by

$$\mathbb{P}_N^H \ell_{\theta,\Lambda}(Y, \Delta, Z) \equiv \mathbb{P}_N^H\{\Delta Z^T\theta + \log \Lambda\{Y\} - \exp(Z^T\theta)\Lambda(Y)\}$$

where $\Lambda(t) = \int \lambda(t)dt$ is the cumulative baseline hazard function, and $\Lambda\{y\}$ is a jump at y. The score for θ and the score operator $B_{\theta,\Lambda} : \mathcal{H} \mapsto L_2(P_{\theta,\Lambda})$ are

$$\dot{\ell}_{\theta,\Lambda}(y, \delta, z) = z\{\delta - e^{\theta^T z}\Lambda(y)\},$$
$$B_{\theta,\Lambda}h(y, \delta, z) = \delta h(y) - e^{\theta^T z}\int_{[0,y]} h d\Lambda,$$

where $\mathcal{H}$ is the unit ball in the space $BV[0, \tau]$ of functions of bounded variation on $[0, \tau]$. The weighted likelihood estimator $(\hat{\theta}_N, \hat{\Lambda}_N)$ solves semiparametric likelihood equations $\mathbb{P}_N^H \dot{\ell}_{\theta,\Lambda} = 0$ and $\mathbb{P}_N^H B_{\theta,\Lambda}(h) = 0$. As in other sampling cases, $\hat{\theta}_N$ is the solution to the weighted partial likelihood equation and $\hat{\Lambda}_N$ is the weighted Breslow estimator (see e.g. Breslow and Wellner, 2007). Conditions for Theorem 2.1 can be verified along the same line as in Saegusa and Wellner (2013) by replacing their empirical process by our weighted empirical process with the help of results in Saegusa (2019). Asymptotic normality of $\hat{\theta}_N$ follows from Theorem 2.1 with asymptotic variance based on the efficient score given by

$$\ell_0^*(y, \delta, z) = \delta(z - (M_1/M_0)(y)) - e^{\theta_0^T z}\int_{[0,y]} (z - (M_1/M_0)(t))\, d\Lambda_0(t),$$

and the efficient information given by

$$I_0 = E\left[\left(\ell_0^*\right)^{\otimes 2}\right] = Ee^{\theta_0^T Z}\int_0^\tau (Z - (M_1/M_0)(y))^{\otimes 2} P(Y \ge y|Z) d\Lambda_0(y),$$

for θ where $M_k(s) = P_{\theta_0,\Lambda_0}[Z^k e^{\theta_0^T Z} I(Y \ge s)]$, $k = 0, 1$. Estimating I_0^{-1} by the plug-in estimator is generally impossible unless covariate vector Z is discrete since it involves the estimation of conditional probability of Y given Z.

For simulation, failure time T was generated from the Cox model with two independent covariates $Z_1 \sim$ Bernoulli(.5) and $Z_2 \sim N(0, 1)$ and the hazard function from Weibull(α, β), $\alpha = .2$, $\beta \in \{0.5, 1, 5\}$ at the baseline. The failure time is subject to censoring by C. The conditional distribution of C given $Z = (Z_1, Z_2)$ is the exponential distribution with rate parameter $c|Z_1 + Z_2|$ where c was chosen to yield approximately 40% and 85% censoring. The regression coefficients are $\theta = (\theta_1, \theta_2)$ with $\theta_1 = \theta_2 = \log(2)$. For data integration, we consider the situation of combining a disease registry with survey data. At the time of sampling, observed variables $V = (\tilde{X}, U)$ are failure status $\tilde{X} = \Delta$ and mailing address U. Data sources are $\mathcal{V}^{(1)} = \{V : \Delta = 1\}$ (disease registry) and $\mathcal{V}^{(2)} = \mathcal{V}$ (survey) with sampling fractions 100% and 20%. For selected subjects, information on $X = (Y, \Delta, Z)$ is available. We generated 2000 data sets with sample sizes $N = 500$, 1000 respectively, and for each data set we generated 2000 bootstrap samples to estimate design variance. The average sample sizes and duplications are shown in Table 1.

Table 1
Sample sizes and the number of duplications based on 1000 simulated data sets. "Censoring" in the first column indicates the censoring proportions and "Dup" in the last column indicates the number of duplications in the final sample.

Censoring	(α, β)	N	$N^{(1)}$	$N^{(2)}$	$n^{(1)}$	$n^{(2)}$	Dup
40%	(.2, .5)	500	306	500	306	100	61
		1000	611	1000	611	200	122
	(.2, 1)	500	308	500	308	100	62
		1000	613	1000	613	200	122
	(.2, 5)	500	300	500	300	100	60
		1000	602	1000	602	200	120
85%	(.2, .5)	500	76	500	76	100	16
		1000	152	1000	152	200	30
	(.2, 1)	500	75	500	75	100	15
		1000	150	1000	150	200	30
	(.2, 5)	500	75	500	75	100	15
		1000	152	1000	152	200	31

Table 2
Results of Monte Carlo simulations for the proposed estimator in data integration and the MLE in the i.i.d. @ setting when the censoring proportion is approximately 40%. SD stands for a Monte Carlo standard deviation, and SEE stands for an average of estimators of standard error, SEE1 stands for an average of estimators of standard error $\Sigma_1^{1/2}$, and SEE2 stands for an average of estimators of standard error $\Sigma_2^{1/2}$ over simulated data sets.

(α, β)		(.2, .5)		(.2, 1)		(.2, 5)	
N		500	1000	500	1000	500	1000
Complete data (MLE)							
θ_1	Bias	.001	.002	.006	.005	.002	.003
	SD	.119	.081	.127	.086	.122	.089
θ_2	Bias	.002	.003	.001	.000	.004	.003
	SD	.067	.051	.075	.049	.078	.057
Data integration							
θ_1	Bias	.001	.003	.010	.013	.002	.003
	SD	.199	.144	.196	.138	.145	.105
	SEE	.196	.142	.178	.129	.140	.100
	SEE1	.120	.085	.121	.085	.124	.087
	SEE2	.155	.114	.130	.097	.060	.047
θ_2	Bias	.016	.003	.016	.005	.005	.003
	SD	.113	.079	.111	.077	.097	.071
	SEE	.103	.076	.104	.075	.091	.065
	SEE1	.069	.049	.073	.051	.079	.055
	SEE2	.077	.058	.073	.055	.041	.033

The Monte Carlo sample bias and standard deviation of the proposed estimator are reported in Table 2 (40% censoring) and Table 3 (85% censoring). The results show that bias in our estimator is small. The proposed variance estimator yielded estimates of the standard errors close to Monte Carlo sample standard deviations of the weighted likelihood estimators in all scenarios. The estimates of $\Sigma_1^{1/2}$ by numerical differentiation of the weighted likelihood are close to the Monte Carlo sample standard deviations of the MLE, which justifies the validity of separate variance estimation based on the decomposition of Σ. The MLE based on the full cohort is more efficient than our proposed estimator as expected. Note that when the MLE was computed based on $N = 1000$ observations, our proposed estimator used information only from selected items of size about 320 on average when the censoring proportion is about 85%. As expected, efficiency of variance estimation is better under moderate censoring than under heavy censoring.

5. Data analysis

We illustrate our methodology with the national Wilms tumor study (NWTS) (D'Angio et al., 1989). Wilms tumor is a rare kidney cancer for children. The NWTS cohort consisted of 3915 patients with Wilms tumor diagnosed during 1980–1994. The patients were followed until the disease progression or death. Baseline covariates are age, stage of cancer, tumor diameter, and histology. The event of interest is disease progression. This data set contains information of all subjects, and was used to compare different hypothetical designs in Breslow and Chatterjee (1999). We compare the proposed weighted likelihood estimator in a data integration setting described below with the MLE based on the entire cohort in the i.i.d. setting. Because of the efficiency of the MLE (see Theorem 2.1 and discussion on Σ that follows), the result of

Table 3
Results of Monte Carlo simulations for the proposed estimator in data integration and the MLE in the i.i.d. @ setting when the censoring proportion is approximately 85%. SD stands for a Monte Carlo standard deviation, SEE stands for an average of estimators of standard error, SEE1 stands for an average of estimators of standard error $\Sigma_1^{1/2}$, and SEE2 stands for an average of estimators of standard error $\Sigma_2^{1/2}$ over simulated data sets.

(α, β)		(.2, .5)		(.2, 1)		(.2, 5)	
N		500	1000	500	1000	500	1000
Complete data (MLE)							
θ_1	Bias	.009	.008	.006	.006	.014	.003
	SD	.246	.172	.251	.179	.414	.273
θ_2	Bias	.003	.003	.003	.001	.001	.000
	SD	.138	.098	.170	.121	.379	.236
Data integration							
θ_1	Bias	.009	.004	.040	.004	.051	.032
	SD	.335	.234	.386	.252	.526	.370
	SEE	.334	.233	.356	.250	.490	.346
	SEE1	.248	.173	.266	.184	.407	.276
	SEE2	.222	.155	.234	.168	.254	.201
θ_2	Bias	.024	.019	.032	.010	.031	.037
	SD	.182	.137	.241	.155	.477	.342
	SEE	.190	.133	.228	.159	.444	.312
	SEE1	.141	.099	.176	.122	.364	.245
	SEE2	.126	.089	.144	.101	.238	.185

Table 4
Point estimates, estimated standard errors, and P-value in the analysis of the NWTS study.

	Full cohort			Data integration		
	$\hat{\theta}$	SE	P-value	$\hat{\theta}$	SE	P-value
Histology	1.357	0.087	<0.001	1.378	0.201	<0.001
Age	0.061	0.015	<0.001	0.104	0.030	<0.001
Stage (III/IV)	1.433	0.257	<0.001	1.542	0.559	0.006
Tumor	0.060	0.015	<0.001	0.041	0.029	0.157
Stage × Tumor	−0.080	0.021	<0.001	−0.075	0.044	0.088

the MLE is considered to be a gold standard. The purpose of data analysis in this section is to illustrate how reasonable the result from our proposed variance estimation is in relation to the gold standard.

We consider the following data integration setting. Two data sources are deceased subjects and the entire cohort with sampling fractions 100%, and 10% respectively resulting in selecting 444 and 392 subjects with 45 duplications. Our goal is to identify predictors of the relapse of Wilms tumor using the Cox proportional hazards model. Table 4 summarizes the result of the proposed method. Our proposed estimator was all close to the MLE. All coefficients were significantly different from 0 in the analysis of the full cohort but analyses of merged data failed to report significance for tumor diameter and its interaction with the stage of cancer due to smaller sample size and complexity of sampling design. This difference reflects the efficiency of the MLE over the proposed weighted likelihood estimator in the sense that the efficient estimator can detect a small effect of covariates on the cancer relapse. On the other hand, the proposed estimator did not find any significant result that the efficient MLE failed to find. These results suggest that our proposed variance estimation method correctly quantifies uncertainty in the analysis of merged data from overlapping sources.

6. Proofs

In this section, we collect proofs of Theorems 2.1–3.2 and auxiliary results. For a function $f_h(x)$ at a fixed x parametrized by $h \in H$, we write $\|f\|_H = \sup_{h\in H} |f_h(x)|$. We first prove the following theorem. Theorem 2.1 is its immediate corollary (see proof of Theorem 2.1 for more details). The result is the extension of Theorem 25.90 of van der Vaart (1998) in the i.i.d. setting to data integration.

Theorem 6.1. *For each θ in a subset of $(\Theta, \|\cdot\|)$ of a normed space and every h in an arbitrary set H, let $x \mapsto \psi_{\theta,h}(x)$ be a measurable function such that the class $\{\psi_{\theta,h} : \|\theta - \theta_0\| < \delta, h \in H\}$ is P_0-Donsker for some $\delta > 0$ with a finite envelope function. Assume that as a map into $\ell^\infty(H)$, the map $\theta \mapsto P_0\psi_{\theta,\cdot}$ is Fréchet differentiable at a zero θ_0, with a derivative $V : \mathrm{lin}\Theta \mapsto \ell^\infty(H)$ that has a continuous inverse on its range. Furthermore, assume that $\|P_0(\psi_{\theta,h} - \psi_{\theta_0,h})^2\|_H \to 0$ as $\theta \to \theta_0$. If $\|\mathbb{P}_N^H \psi_{\hat{\theta}_N}\|_H = o_P(N^{-1/2})$ and $\hat{\theta}_N \to_P \theta_0$, then*

$$\sqrt{N}V(\hat{\theta}_N - \theta_0) = -\mathbb{G}_N^H \psi_{\theta_0} + o_P(1).$$

Proof. We prove $\sqrt{N}\|\hat{\theta}_N - \theta_0\| = O_P(1)$ assuming

$$\mathbb{G}_N^H(\psi_{\hat{\theta}_n,h} - \psi_{\theta_0,h}) = o_P(1). \tag{10}$$

Because $\mathbb{P}_N^H \psi_{\hat{\theta}_N,h} = o_P(N^{-1/2})$ and $P_0\psi_{\theta_0,h} = 0$, it follows that

$$\begin{aligned}\sqrt{N}(P_0\psi_{\hat{\theta}_N,h} - P_0\psi_{\theta_0,h}) &= \sqrt{N}(P_0\psi_{\hat{\theta}_N} - \mathbb{P}_N^H\psi_{\hat{\theta}_N}) + o_P(1)\\ &= -\sqrt{N}(\mathbb{P}_N^H\psi_{\theta_0,h} - P_0\psi_{\theta_0,h}) + o_P(1).\end{aligned} \tag{11}$$

In the last step we used the assumption (10). Since the continuous invertibility of V at θ_0 implies that there is some constant $c > 0$ such that $(c + o_P(1))\|\hat{\theta}_N - \theta_0\| \le \|P_0\psi_{\hat{\theta}_N,h} - P_0\psi_{\theta_0,h}\|_{\mathcal{H}}$, we have

$$(c + o_P(1))\sqrt{N}\|\hat{\theta}_N - \theta_0\| \le \|\sqrt{N}(P_0\psi_{\hat{\theta}_N} - P_0\psi_{\theta_0})\|_{\mathcal{H}} \le \|\mathbb{G}_N^H\psi_{\theta_0}\|_{\mathcal{H}} + o_P(1).$$

Since $\|\mathbb{G}_N^H\psi_{\theta_0}\|_{\mathcal{H}} = O_P(1)$ by assumption and Theorem 3.2 of Saegusa (2019), the claim $\sqrt{N}\|\hat{\theta}_N - \theta_0\| = O_P(1)$ follows.

Now, the desired result follows from the differentiability of $\theta \mapsto P_0\psi_\theta$ and $\sqrt{N}$-consistency of $\hat{\theta}_N$ that (11) becomes

$$\sqrt{N}V(\hat{\theta}_N - \theta_0) = -\mathbb{G}_N^H\psi_{\theta_0} + o_P(1).$$

Now we prove $\mathbb{G}_N^H(\psi_{\hat{\theta}_n,h} - \psi_{\theta_0,h}) = o_P(1)$. Since $\hat{\theta}_N \to_P \theta_0$, we can assume without loss of generality that $\hat{\theta}_N$ takes its values in $\Theta_\delta = \{\theta \in \Theta : \|\theta - \theta_0\| < \delta\}$. For an element of $\ell^\infty(\Theta_\delta \times H)$, we write z with a value $z(\theta, h)$ evaluated at $(\theta, h) \in \Theta_\delta \times H$. Now define the map $g : \ell^\infty(\Theta_\delta \times H) \times \Theta_\delta \mapsto \ell^\infty(H)$ by

$$g(z, \theta)(h) \mapsto z(\theta, h) - z(\theta_0, h).$$

This map is continuous at every point (z, θ_0) such that $\|z(\theta, h) - z(\theta_0, h)\|_H \to 0$ at all sample paths z of the limit process Z of the stochastic process $Z_n(\theta, h) = \mathbb{G}_N^H\psi_{\theta,h}$ indexed by $\Theta_\delta \times H$. Because the index set for $\mathbb{G}_N^H$ as a process is P-Donsker by assumption, Theorem 3.2 of Saegusa (2019) implies that Z_n weakly converges to a tight Gaussian process Z which is a linear combination of independent Brownian bridge processes. Because the covariance function of the limit process $\mathbb{G}$ of $\mathbb{G}_N^H$ is bounded above by the covariance function of the limit process of the standard empirical process $\mathbb{G}_N$, the limit process Z has continuous sample paths with respect to the semimetric d given by

$$d^2((\theta_1, h_1), (\theta_2, h_2)) = P_0(\psi_{\theta_1,h_1} - \psi_{\theta_2,h_2})^2.$$

Since $\sup_h d((\theta, h), (\theta_0, h)) \to 0$ as $\theta \to \theta_0$ by assumption, we can conclude that g is continuous at almost all sample paths z of Z.

By Slutsky's theorem, $(Z_n, \hat{\theta}_n) = (\mathbb{G}_N^H\psi_\theta, \hat{\theta}_n)$ converges in distribution to $(\mathbb{G}, \theta_0)$. By the continuous mapping theorem,

$$\mathbb{G}_N^H(\psi_{\hat{\theta}_n,h} - \psi_{\theta_0,h}) = g(Z_n, \hat{\theta}_n) \to_d g(Z, \theta_0) = 0. \quad \square$$

Now we apply Theorem 6.1 to the proposed semiparametric weighted likelihood estimator. In Theorem 6.1, the parameter of interest is θ and the derivative V of the map $\theta \mapsto P_0\psi_\theta$ is abstract. In Theorem 2.1, the parameter of interest is (θ, η) and the derivative map $\dot{\Psi}_0$ can be partitioned and has explicit expressions.

Proof of Theorem 2.1. Theorem 2.1 is a corollary to Theorem 6.1. To see this, replace θ, ψ_θ, and V in Theorem 6.1 by (θ, η), $(\dot{\ell}_{\theta,\eta}, B_{\theta,\eta} - P_{\theta,\eta}B_{\theta,\eta})$, and $\dot{\Psi}$, respectively. The corresponding conditions regarding the set $\{\psi_\theta : \|\theta - \theta_0\| < \delta, h \in H\}$ and the quantity $\|P_0(\psi_{\theta,h} - \psi_{\theta_0,h})^2\|_H$ can be found in Condition 1. The condition $\|\mathbb{P}_N^H\psi_{\hat{\theta}_N}\|_H = o_P(N^{-1/2})$ corresponds to the semiparametric likelihood equations (2). The corresponding condition regarding V can be found in Condition 2. The conclusion of Theorem 6.1 in the context of Theorem 2.1 is that

$$\dot{\Psi}_0\sqrt{N}(\hat{\theta}_N - \theta_0, \hat{\eta}_N - \eta_0) = \mathbb{G}_N^H(\dot{\ell}_{\theta_0,\eta_0}, B_{\theta_0,\eta_0}) + o_P(1).$$

Write $\int f d\eta = \eta f$ for a measure η and $I_0 = P_0\dot{\ell}_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}^T$. Using the expression of $\dot{\Psi}$ in Condition 2 the above expression can be written as

$$\begin{aligned}&-I_0\sqrt{N}(\hat{\theta}_N - \theta_0) - \sqrt{N}(\hat{\eta}_N - \eta_0)B_{\theta_0,\eta_0}^*\dot{\ell}_{\theta_0,\eta_0} = -\mathbb{G}_N^H\dot{\ell}_{\theta_0,\eta_0} + o_P(1),\\ &-P_0\{(B_{\theta_0,\eta_0}h)\dot{\ell}_{\theta_0,\eta_0}^T\}\sqrt{N}(\hat{\theta}_N - \theta_0) - \sqrt{N}(\hat{\eta}_N - \eta_0)B_{\theta_0,\eta_0}^*B_{\theta_0,\eta_0}h\\ &\quad = -\mathbb{G}_N^H B_{\theta_0,\eta_0}h + o_P(1).\end{aligned}$$

Choose $h = (B_{\theta_0,\eta_0}^*B_{\theta_0,\eta_0})^{-1}B_{\theta_0,\eta_0}^*\dot{\ell}_{\theta_0,\eta_0}$ in the second equation to obtain

$$\begin{aligned}&-P_0(B_{\theta_0,\eta_0}(B_{\theta_0,\eta_0}^*B_{\theta_0,\eta_0})^{-1}B_{\theta_0,\eta_0}^*\dot{\ell}_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}^T)\sqrt{N}(\hat{\theta}_N - \theta_0)\\ &-\sqrt{N}(\hat{\eta}_N - \eta_0)B_{\theta_0,\eta_0}^*\dot{\ell}_{\theta_0,\eta_0}\\ &= -\mathbb{G}_N^H B_{\theta_0,\eta_0}(B_{\theta_0,\eta_0}^*B_{\theta_0,\eta_0})^{-1}B_{\theta_0,\eta_0}^*\dot{\ell}_{\theta_0,\eta_0}o_P(1).\end{aligned}$$

Recall the definitions of the efficient information $\tilde{I}_0$ for θ and the efficient influence function $\tilde{\ell}_0$ for θ. Subtract the second equation from the first and matrix inversion of $\tilde{I}_0$ yield

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbb{G}_N^H \tilde{\ell}_0 + o_P(1).$$

Apply Theorem 3.2 of Saegusa (2019) to obtain the desired result that $\mathbb{G}_N^H \tilde{\ell}_0$ converges in distribution to a zero mean normal random vector with covariance Σ. □

Proof of Theorem 3.1. Let $\overline{\theta}_N = \hat{\theta}_N + h_N v_N$. We compute lower and upper bounds of

$$\mathbb{M}_N(\overline{\theta}_N) - \mathbb{M}_N(\hat{\theta}_N) = \mathbb{P}_N^H \ell_{\overline{\theta}_N, \hat{\eta}_{\overline{\theta}_N, N}} - \mathbb{P}_N^H \ell_{\hat{\theta}_N, \hat{\eta}_{\hat{\theta}_N, N}}.$$

For the lower bound, we have by the definitions of $\hat{\eta}_\theta$ and $t \mapsto \eta_\theta(\gamma)$ and (4) that

$$\begin{aligned}\mathbb{M}_N(\overline{\theta}_N) - \mathbb{M}_N(\hat{\theta}_N) &\geq \mathbb{P}_N^H \ell_{\overline{\theta}_N, \eta_{\overline{\theta}_N}(\hat{\gamma}_{\hat{\theta}_N, N})} - \mathbb{P}_N^H \ell_{\hat{\theta}_N, \eta_{\hat{\theta}_N}(\hat{\gamma}_{\hat{\theta}_N, N})} \\ &= \ell(\overline{\theta}_N, \hat{\gamma}_{\hat{\theta}_N, N}) - \ell(\hat{\theta}_N, \hat{\gamma}_{\hat{\theta}_N, N}).\end{aligned}$$

Since $\ell(t, \gamma)$ is differentiable with respect to t, we expand the lower bound around $\hat{\theta}_N$ to obtain

$$h_N v_N^T \mathbb{P}_N^H \dot{\ell}(\hat{\theta}_N, \hat{\gamma}_{\hat{\theta}_N, N}) + 2^{-1} h_N^2 v_N^T \mathbb{P}_N^H \ddot{\ell}(\tilde{\theta}_N, \hat{\gamma}_{\hat{\theta}_N, N}) v_N,$$

where $\tilde{\theta}_N$ is some convex combination of $\overline{\theta}_N$ and $\hat{\theta}_N$. The first term is zero because the map $t \mapsto \mathbb{P}_N^H \ell_{t, \eta_t(\hat{\gamma}_{\hat{\theta}_N, N})}$ is maximized at $t = \hat{\theta}_N$. The second term becomes $-2^{-1} h_N^2 (v_N^T \tilde{I}_0 v_N + o_P(1))$ by (6).

For the upper bound, we have by (4) that

$$\begin{aligned}\mathbb{M}_N(\overline{\theta}_N) - \mathbb{M}_N(\hat{\theta}_N) &\leq \mathbb{P}_N^H \ell_{\overline{\theta}_N, \eta_{\overline{\theta}_N}(\hat{\gamma}_{\overline{\theta}_N, N})} - \mathbb{P}_N^H \ell_{\hat{\theta}_N, \eta_{\hat{\theta}_N}(\hat{\gamma}_{\overline{\theta}_N, N})} \\ &= \ell(\overline{\theta}_N, \hat{\gamma}_{\overline{\theta}_N, N}) - \ell(\hat{\theta}_N, \hat{\gamma}_{\overline{\theta}_N, N}).\end{aligned}$$

We expand the upper bound around $\overline{\theta}_N$ to obtain

$$h_N v_N^T \mathbb{P}_N^H \dot{\ell}(\overline{\theta}_N, \hat{\gamma}_{\overline{\theta}_N, N}) - 2^{-1} h_N^2 v_N^T \mathbb{P}_N^H \ddot{\ell}(\tilde{\theta}_N, \hat{\gamma}_{\overline{\theta}_N, N}) v_N$$

where $\tilde{\theta}_N$ is some convex combination of $\overline{\theta}_N$ and $\hat{\theta}_N$. The second term is $2^{-1} h_N^2 (v_N^T \tilde{I}_0 v_N + o_P(1))$ by (6). The first term is equal to

$$\begin{aligned}&\frac{h_N}{\sqrt{N}} v_N^T \mathbb{G}_N^H \dot{\ell}(\overline{\theta}_N, \hat{\gamma}_{\overline{\theta}_N, N}) + h_N v_N^T P_0 \dot{\ell}(\overline{\theta}_N, \hat{\gamma}_{\overline{\theta}_N, N}) \\ &= \frac{h_N}{\sqrt{N}} (v_N^T \tilde{I}_0 \sqrt{N}(\hat{\theta}_N - \theta_0) + o_P(1)) \\ &\quad - h_N \left\{ v_N^T \tilde{I}_0 (\overline{\theta}_N - \theta_0) + o_P(\|\overline{\theta}_N - \theta_0\| + N^{-1/2}) \right\},\end{aligned}$$

by the proof of Theorem 2.1, (5) and (7). This is equal to $-h_N^2 (v_N^T \tilde{I}_0 v_N + o_P(1))$ by the assumption of h_N and the definition of $\overline{\theta}_N$.

Combining the upper and lower bounds, we have

$$v_N^T \tilde{I}_0 v_N + o_P(1) \leq -2 \frac{\mathbb{M}_N(\overline{\theta}_N) - \mathbb{M}_N(\hat{\theta}_N)}{h_N^2} \leq v_N^T \tilde{I}_0 v_N + o_P(1).$$

Taking $N \to \infty$, we obtain the desired result. □

Proof of Theorem 3.2. We prove that the limiting process is the same as the limiting process of $\mathbb{G}_N^H$ except the process due to sampling from population. Once we establish this claim, the rest is similar to the proof of Theorem 2.1 by replacing $\mathbb{G}_N^H$ and $\mathbb{P}_N^H$ by $\hat{\mathbb{G}}_N^H$ and $\hat{\mathbb{P}}_N^H$. Define the bootstrap empirical process by $\hat{\mathbb{G}}_N^H = \sqrt{N}(\hat{\mathbb{P}}_N^H - \mathbb{P}_N^H)$. Then we have

$$\begin{aligned}\hat{\mathbb{G}}_N^H &= \frac{1}{\sqrt{N}} \sum_{j=1}^J \sum_{i=1}^N \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \rho^{(j)}(V_i)(B_i^{(j)} - 1)\delta_{X_i} \\ &= \sum_{j=1}^J \sqrt{\frac{N^{(j)}}{N}} \sqrt{\frac{N^{(j)}}{n^{(j)}}} \frac{1}{\sqrt{n^{(j)}}} \sum_{i=1}^N R_i^{(j)}(B_i^{(j)} - 1)\rho^{(j)}(V_i)\delta_{X_i} \\ &\equiv \sum_{j=1}^J \sqrt{\frac{N^{(j)}}{N}} \sqrt{\frac{N^{(j)}}{n^{(j)}}} \hat{\mathbb{G}}_N^{H,(j)}\end{aligned}$$

Conditionally on $(X_i R_i, R_i, V_i)$, the bootstrap process $\hat{\mathbb{G}}_N^{H,(j)}$ for data source j can be viewed as the sampling-without-replacement bootstrap process with sample size $n^{(j)}$ with the index set $\{g(x,v) = \rho^{(j)}(v)f(x) : f \in \mathcal{F}\}$ for some index set $\tilde{\mathcal{F}}^{(j)} \equiv \mathcal{F}$ for the original weighted empirical process $\mathbb{G}_N^H$ (see Saegusa, 2015 for more details). It follows by the exchangeably weighted bootstrap empirical process theory (Præstgaard and Wellner, 1993) that $\hat{\mathbb{G}}_N^{H,(j)}$ weakly converges to $\sqrt{1-p^{(j)}}\mathbb{G}^{(j)}$ in $\ell^\infty(\tilde{\mathcal{F}}^{(j)})$ conditionally on data where $\mathbb{G}^{(j)}$ is the Brownian bridge process with respect to the conditional probability measure given the membership in source j. Thus,

$$\hat{\mathbb{G}}_N^H \rightsquigarrow \sum_{j=1}^{J} \sqrt{v_j}\sqrt{\frac{1-p^{(j)}}{p^{(j)}}}\mathbb{G}^{(j)}$$

conditionally on data. Note that $\mathbb{G}^{(j)}$ are all independent because $B_i^{(j)}$ and $B_i^{(j')}$ are independent given data. The limiting process of $\hat{\mathbb{G}}_N^H$ above is exactly the same as the limiting process $\mathbb{G}_N^H$ except the process due to sampling from population. □

7. Discussion and future work

In this paper, we study semiparametric inference for merged data from multiple sources. We derive the asymptotic distribution of the semiparametric weighted likelihood estimator of the finite dimensional parameter. As in the i.i.d. setting, the asymptotic variance in the proposed estimator may not have a closed form or contains expectations of unknown functions in many semiparametric models. We developed a consistent computational method to estimate asymptotic variance by the numerical derivative of the profile likelihood and the sampling-without-replacement bootstrap. This is the first variance estimation method for data integration when a plug-in variance estimator is not available.

To illustrate our methodology, we studied the Cox proportional hazards model with right censoring for simulation and data analysis. As mentioned in Introduction, there are many other important semiparametric models which finds potential applications in data integration problems. Estimators in these models solve infinite-dimensional optimization problems and the forms of asymptotic variance are expected to be complicated as in the i.i.d. setting. Our proposed variance estimator will continue to be useful for these estimators because our method only requires computation of estimators and weighted likelihood. In addition to methods research, we will develop software to implement our methodology for major semiparametric models to encourage semiparametric data integration in practice. These work will appear elsewhere.

The proposed variance estimation methodology focuses on weighted semiparametric likelihood estimators. Other estimation procedures such as the estimating equations approach and the sieve weighted likelihood approach (Geman and Hwang, 1982; Shen, 1997) are also possible in data integration problems with the help of inverse probability weighting and ρ. For these estimators, our proposed method of estimating the efficient information may not be valid. First, those estimators have asymptotic variance which does not involve the efficient information. Second, variance estimation may not be consistent due to unmet regularity conditions. The sieve weighted likelihood estimator, for example, maximizes the weighted likelihood over smaller parameter spaces. The least favorable submodel is not necessarily contained in the corresponding smaller model space and hence the sieve profile likelihood may not approximate the likelihood based on the least favorable submodel. For estimators other than what we considered in this paper, further methodological development is desired.

Acknowledgments

This research was funded by the National Science Foundation, USA (DMS 2014971). I would like to thank the editor, the associate editor, and the referee for their constructive comments and suggestions.

References

Bickel, P.J., Freedman, D.A., 1984. Asymptotic normality and the bootstrap in stratified sampling. Ann. Statist. 12 (2), 470–482. http://dx.doi.org/10.1214/aos/1176346500.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., Wellner, J.A., 1998. Efficient and Adaptive Estimation for Semiparametric Models. Springer-Verlag, New York, p. xxii+560, Reprint of the 1993 original.

Breslow, N.E., Chatterjee, N., 1999. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. J. R. Stat. Soc. Ser. C. Appl. Stat. 48 (4), 457–468. http://dx.doi.org/10.1111/1467-9876.00165.

Breslow, N.E., Wellner, J.A., 2007. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scand. J. Stat. 34 (1), 86–102.

Brick, J.M., Dipko, S., Presser, S., Tucker, C., Yuan, Y., 2006. Nonresponse bias in a dual frame sample of cell and landline numbers. Publ. Opin. Q. 70 (5), pp. 780–793.

Cervantes, I., Jones, M., Rojas, L., Brick, J., Kurata, J., Grant, D., 2006. A review of the sample design for the california health interview survey. In: Proceedings of the Social Statistics Section. American Statistical Association, pp. 3023–3030.

Chatterjee, N., Chen, Y.-H., Maas, P., Carroll, R.J., 2016. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. J. Amer. Statist. Assoc. 111 (513), 107–117. http://dx.doi.org/10.1080/01621459.2015.1123157.

D'Angio, G.J., Breslow, N., Beckwith, J.B., Evans, A., Baum, H., deLorimier, A., Fernbach, D., Hrabovsky, E., Jones, B., Kelalis, P., 1989. Treatment of Wilms' tumor. Results of the Third National Wilms' Tumor Study. Cancer 64 (2), 349–360.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Statist. 7 (1), 1–26.

Geman, S., Hwang, C.-R., 1982. Nonparametric maximum likelihood estimation by the method of sieves. Ann. Statist. 10 (2), 401–414.

Gross, S., 1980. Median estimation in sample surveys. In: Proceedings of the Section on Survey Research Methods, pp. 181–184.
Hartley, H.O., 1962. Multiple frame surveys. In: Proceedings of the Social Statistics Section. American Statistical Association, pp. 203–206.
Hartley, H.O., 1974. Multiple frame methodology and selected applications. Sankhyā C 36, 99–118.
Hu, S.S., Balluz, L., Battaglia, M.P., Frankel, M.R., 2011. Improving public health surveillance using a dual-frame survey of landline and cell phone numbers. Am. J. Epidemiol. 173 (6), 703–711. http://dx.doi.org/10.1093/aje/kwq442.
Huang, J., 1996. Efficient estimation for the proportional hazards model with interval censoring. Ann. Statist. 24 (2), 540–568.
Keiding, N., Louis, T.A., 2016. Perils and potentials of self-selected entry to epidemiological studies and surveys. J. Roy. Statist. Soc. Ser. A 179 (2), 319–376. http://dx.doi.org/10.1111/rssa.12136.
Kosorok, M.R., 2008. Introduction To Empirical Processes and Semiparametric Inference. In: Springer Series in Statistics, Springer, New York, p. xiv+483. http://dx.doi.org/10.1007/978-0-387-74978-5.
Metcalf, P., Scott, A., 2009. Using multiple frames in health surveys. Stat. Med. 28 (10), 1512–1523. http://dx.doi.org/10.1002/sim.3566.
Murphy, S.A., 1995. Asymptotic theory for the frailty model. Ann. Statist. 23 (1), 182–198.
Murphy, S.A., Rossini, A.J., van der Vaart, A.W., 1997. Maximum likelihood estimation in the proportional odds model. J. Amer. Statist. Assoc. 92 (439), 968–976. http://dx.doi.org/10.2307/2965560.
Murphy, S.A., van der Vaart, A.W., 1996. Likelihood inference in the errors-in-variables model. J. Multivariate Anal. 59 (1), 81–108. http://dx.doi.org/10.1006/jmva.1996.0055.
Murphy, S.A., van der Vaart, A.W., 1999. Observed information in semi-parametric models. Bernoulli 5 (3), 381–412.
Murphy, S.A., van der Vaart, A.W., 2000. On profile likelihood. J. Amer. Statist. Assoc. 95 (450), 449–485, With comments and a rejoinder by the authors, http://dx.doi.org/10.2307/2669386, https://doi.org/10.2307/2669386.
Nan, B., Kalbfleisch, J.D., Yu, M., 2009. Asymptotic theory for the semiparametric accelerated failure time model with missing data. Ann. Statist. 37 (5A), 2351–2376. http://dx.doi.org/10.1214/08-AOS657, URL https://doi-org.proxy-um.researchport.umd.edu/10.1214/08-AOS657.
Parner, E., 1998. Asymptotic theory for the correlated gamma-frailty model. Ann. Statist. 26 (1), 183–214. http://dx.doi.org/10.1214/aos/1030563982.
Præstgaard, J., Wellner, J.A., 1993. Exchangeably weighted bootstraps of the general empirical process. Ann. Probab. 21 (4), 2053–2086.
Quenouille, M.H., 1949. Problems in plane sampling. Ann. Math. Statist. 20, 355–375. http://dx.doi.org/10.1214/aoms/1177729989.
Saegusa, T., 2015. Variance estimation under two-phase sampling. Scand. J. Stat. 42 (4), 1078–1091. http://dx.doi.org/10.1111/sjos.12152.
Saegusa, T., 2019. Large sample theory for merged data from multiple sources. Ann. Statist. 47 (3), 1585–1615. http://dx.doi.org/10.1214/18-AOS1727.
Saegusa, T., Wellner, J.A., 2013. Weighted likelihood estimation under two-phase sampling. Ann. Statist. 41 (1), 269–295. http://dx.doi.org/10.1214/12-AOS1073.
Shao, J., Tu, D.S., 1995. The Jackknife and Bootstrap. In: Springer Series in Statistics, Springer-Verlag, New York, p. xviii+516. http://dx.doi.org/10.1007/978-1-4612-0795-5.
Shen, X., 1997. On methods of sieves and penalization. Ann. Statist. 25 (6), 2555–2591. http://dx.doi.org/10.1214/aos/1030741085.
Tukey, J.W., 1958. Bias and confidence in not quite large samples (abstract). Ann. Math. Statist. 29 (2), 614.
van der Vaart, A.W., 1998. Asymptotic Statistics. In: Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3, Cambridge University Press, Cambridge, p. xvi+443.
Zhang, Y., Hua, L., Huang, J., 2010. A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. Scand. J. Stat. 37 (2), 338–354. http://dx.doi.org/10.1111/j.1467-9469.2009.00680.x.