The Annals of Statistics
2019, Vol. 47, No. 3, 1585–1615
https://doi.org/10.1214/18-AOS1727
© Institute of Mathematical Statistics, 2019

LARGE SAMPLE THEORY FOR MERGED DATA FROM MULTIPLE SOURCES

BY TAKUMI SAEGUSA

University of Maryland

We develop large sample theory for merged data from multiple sources. Main statistical issues treated in this paper are (1) the same unit potentially appears in multiple datasets from overlapping data sources, (2) duplicated items are not identified and (3) a sample from the same data source is dependent due to sampling without replacement. We propose and study a new weighted empirical process and extend empirical process theory to a dependent and biased sample with duplication. Specifically, we establish the uniform law of large numbers and uniform central limit theorem over a class of functions along with several empirical process results under conditions identical to those in the i.i.d. setting. As applications, we study infinite-dimensional M-estimation and develop its consistency, rates of convergence and asymptotic normality. Our theoretical results are illustrated with simulation studies and a real data example.

1. Introduction. Many organizations nowadays collect massive datasets from various sources including online surveys, business transactions, social media and scientific research. In contrast to well-controlled small data, the representativeness of these datasets often critically depends on technology for data collection. A promising remedy to reduce potential selection bias is to merge multiple samples with different coverages. Data integration problems, however, have not been fully studied in view of basic limit theorems such as the law of large numbers (LLN) and the central limit theorem (CLT). The main statistical challenges we focus on here are (1) potential duplicated selection from overlapping sources of different sizes, (2) the lack of identification of duplicated items across datasets and (3) dependence among observations in each source induced by sampling without replacement. Because large parts of statistical theory rely on the assumption that observations are independent and identically distributed (i.i.d.), the analysis of merged data from multiple sources requires a novel approach in theory and methods.

The basic setting considered in this paper is described below. This is also illustrated in Figure 1.

• Our interest lies in a statistical model $\mathcal{P}$ for a vector of variables X taking values in a measurable space $(\mathcal{X}, \mathcal{A})$. Suppose $X \sim P_0 \in \mathcal{P}$.

Received February 2017; revised May 2018.
MSC2010 subject classifications. Primary 62E20; secondary 62G20, 62D99, 62N01.
Key words and phrases. Calibration, data integration, empirical process, nonregular, sampling without replacement, semiparametric model.

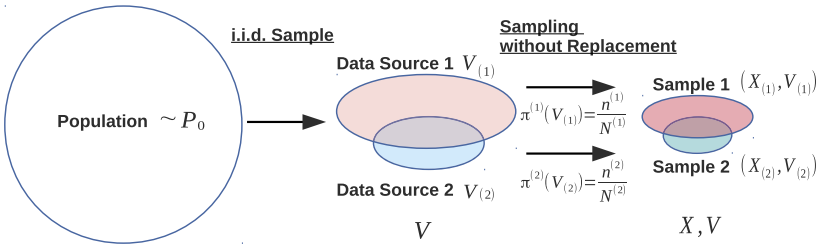

FIG. 1. *Sampling scheme for merged data from multiple sources with* $J = 2$.

• Let $V = (\check{X}, U) \in \mathcal{V}$ where $\check{X}$ is a coarsening of X and U is a vector of auxiliary variables that do not contain information about the model $\mathcal{P}$. The space $\mathcal{V}$ consists of J overlapping "(population) data sources" $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(J)}$ with $\bigcup_{j=1}^{J} \mathcal{V}^{(j)} = \mathcal{V}$ and $\mathcal{V}^{(j)} \cap \mathcal{V}^{(j')} \neq \varnothing$ for some (j, j'). Variables V determine source membership.

• For data collection, a large sample is drawn from a population: let $V_1, \ldots, V_N$ be i.i.d. as V. Unit i belongs to source j if $V_i \in \mathcal{V}^{(j)}$. Sample size in source j is $N^{(j)} = \#\{i \leq N : V_i \in \mathcal{V}^{(j)}\}$.

• Next, a random sample of size $n^{(j)}$ is drawn without replacement from source j with sampling probability $\pi^{(j)}(V_i) = (n^{(j)}/N^{(j)})I\{V_i \in \mathcal{V}^{(j)}\}$. For selected units, we observe X. We repeat the same process for all sources. As $n^{(j)} \leq N^{(j)}$ holds, $n^{(j)}$ and $\pi^{(j)}(\cdot)$ are a random variable and a random function, respectively.

• Finally, multiple datasets from different sources are combined. Our proposed estimation method estimates the parameters of the model $\mathcal{P}$.

The two-stage formulation is crucial in describing duplicated selection. A large sample is drawn from a population (*sampling from population*), and units are classified into one or more (sample) data sources. Next, subsamples are drawn without replacement from each data source (*finite-population sampling*) to generate multiple datasets. The sample at the first stage serves as a finite population to allow for repeated selection of the same units.

Information that statisticians have at their disposal is the X- and V-values of the selected items from different sources, membership information on (other) data sources to which selected items belong, and the realizations of $N^{(j)}$ and $n^{(j)}$. A special case where V-values are also available for non-sampled items is treated in Section 4.

Our framework covers a number of applications. Typical examples are opinion polls [9], public health surveillance [30] and health interview surveys [11] where data sources are lists of cell- and landline phone users. Duplicated records in databases are important issues in business operations [28]. Scientific research has considered combining face-to-face, telephone and online surveys [15, 17]. Our setting also covers the situation where one data source is entirely contained in

another. This case is highly useful for studying rare disease and rare exposure represented as smaller data sources [33, 35]. Applications include the synthesis of existing clinical and epidemiological studies with surveys, disease registries and healthcare databases [12, 36, 45].

Despite scientific and financial benefits of data integration, many important models have never been studied in our setting due to the lack of probabilistic tools to study a dependent and biased sample with duplication. We address this issue by extending empirical process theory with applications to infinite-dimensional M-estimation in mind. This theory provides essential tools for the analysis of semi- and nonparametric inference (see, e.g., [38, 58]). It originated in the study of the uniform law of large numbers (U-LLN) and the uniform central limit theorem (U-CLT) in the i.i.d. setting [10, 19, 20, 22, 23]. The i.i.d. assumption has been relaxed in several directions including triangular arrays [60, 61], martingale difference [40], Markov chains [2] and stationary processes [3]. The study of dependent empirical processes arising from complex sampling was initiated by [7] for stratified samples followed by [52]. Beyond stratified samples, [4] and [5] studied the U-CLT for rejective sampling and single stage sampling, respectively.

Our sampling scheme is markedly different from those in the above literature in important ways. A basic technique to analyze dependent empirical processes is to find a hidden (nearly) independent structure as seen in [4] that utilized similarity between independent Poisson sampling and rejective sampling. This method needs a simple dependence structure but our merged data have complex multitiered dependence: First, items within the same source are dependent due to sampling without replacement. Second, items across overlapping sources are dependent because they are potentially identical. Previous studies focused on dependence within a sample but our theory addresses dependence within and between samples at the same time. Another difference is that simple inverse probability weighting adopted in [4, 5, 7] is not valid in our setting. This technique corrects selection bias from data sources but does not account for bias from duplicated selection.

We build large sample theory on a novel weighted empirical process that integrates information from multiple sources. Our main contribution is the U-LLN and U-CLT over a class of functions. We only assume that an index set is Glivenko–Cantelli or Donsker as in [7, 52]. This implies that if the U-LLN or the U-CLT holds for the i.i.d. sample, the corresponding results hold for merged data without additional conditions. This formulation is of practical importance because fair comparison can be made between previous scientific conclusions from i.i.d. samples and the ones from the analysis of merged data without worrying about differences in assumptions. This generality makes a contrast with [4] that assumes the uniform entropy condition and [5] that assumes a priori the existence of the finite-dimensional CLT.

Another contribution is theory of infinite-dimensional M-estimation for merged data. Previous research tended to focus on the U-CLT with limited applications as

a result (e.g., statistical functionals in [4, 5]), but the U-LLN and maximal inequalities are essential to obtain consistency and rates of convergence for M-estimators. We obtain a set of empirical process tools beyond the U-CLT, and derive consistency, rates of convergence and asymptotic normality of our estimators. We obtain optimal calibration [16, 48] and optimal weights in our weighted empirical process that improve efficiency of our estimators. We study several examples including the Cox proportional hazards models [13] and illustrate the finite sample performance of our methods through numerical studies in several different scenarios.

Our theory can be viewed as a nontrivial extension of [7, 52] for stratified samples to overlapping "strata." In stratified sampling, the i.i.d. sample from population is stratified and finite population sampling is carried out in each stratum. One may consider our sampling scheme as "stratified sampling" with nonnegligible intersections among strata. The approach of [7, 52] is, however, not applicable to our setting due to issues of multitiered dependence and inverse probability weighting discussed above. In particular, their proof exploited the disjoint nature of strata and reduced weak convergence to multiple convergence within strata. This method addresses dependence within strata but does not cover dependence across "strata" arising from duplicated selection (see Section 3 for details). Note that our framework is more general than previously studied sampling designs including stratified sampling in that it accommodates those designs in place of finite population sampling. In Section D of the Supplementary Material [51], we treat stratified sampling at the second stage of sampling in the data integration context.

The rest of the paper is organized as follows. In Section 2, we introduce our weighted empirical process and discuss more on our sampling framework. We present the U-LLN and several variants of U-CLTs in Section 3. Calibration methods are treated in Section 4. We study infinite-dimensional M- and Z-estimation and their applications in Section 5. Finite sample properties of proposed methods are illustrated in numerical studies in Section 6. Section 7 discusses differences between our framework and those in sampling theory. All proofs and additional simulation are given in the Supplementary Material [51].

2. Sampling and empirical process. We review basic settings and introduce our weighted empirical process.

2.1. *Sampling.* Let $R_i^{(j)} \in \{0, 1\}$ be a sampling indicator from source j. Simple random sampling from each source is carried out independently. Thus, sampling indicators $(R_1^{(j)}, \ldots, R_N^{(j)})$ and $(R_1^{(j')}, \ldots, R_N^{(j')})$ with $j \neq j'$ are conditionally independent given $V_1, \ldots, V_N$. However, sampling indicators within the same source are not independent but are only exchangeable due to sampling without replacement. The unit that does not belong to source j (i.e., $V_i \notin \mathcal{V}^{(j)}$) automatically has $R_i^{(j)} = 0$. Throughout we denote inverse probability weighting by $R_i^{(j)}/\pi^{(j)}(V_i)$ with convention $0/0 = 0$.

To enumerate units within a data source, we write, for example, $X_{(j),i}$ to mean the observation of X for the unit i in source j with index i going from 1 through $N^{(j)}$ [see, e.g., (3.1)]. The limits of sampling probabilities are $\lim_{N\to\infty}\pi^{(j)}(v) = p^{(j)}I\{v \in \mathcal{V}^{(j)}\}$ where $p^{(j)} \geq c > 0$ for some constant c. We assume N is known. In Section F of the Supplementary Material [51], we consider the case of unknown N which may be the case in practice. For additional notation, let $W = (X, U) \in \mathcal{X} \times \mathcal{U} \equiv \mathcal{W}$ with $W \sim \tilde{P}_0$. The conditional measure given membership in source j is denoted as $P_0^{(j)}$, that is, for measurable $A \subset \mathcal{W}$, $P_0^{(j)}(A) = \tilde{P}_0(A \cap \mathcal{V}^{(j)})/\nu^{(j)}$ where $\nu^{(j)} \equiv \tilde{P}_0(V \in \mathcal{V}^{(j)})$ is membership probability in source j. The conditional probability measure for $R_i^{(j)}$ given $N^{(j)}$, $i = 1, \ldots, N$, $j = 1, \ldots, J$, is denoted as $P_{R,N}$. The probability measure P^∞ is defined such that its projection of the first N coordinates is $\tilde{P}_0^N \times P_{R,N}$.

2.2. *Assumption of unidentified duplication.* Duplicated items are not identified in our setting, which reflects the lack of communication between sampling procedures. Instead, we assume that we can identify additional data source membership of selected items by checking their V. This assumption is not too restrictive. For example, telephone surveys can ask an additional question whether to own both landline and cell phones. When medical studies are merged, comparison of inclusion and exclusion criteria suffices. Identifying duplication, on the other hand, produces unavoidable errors. Important identifiers such as names, addresses and social security numbers are usually not disclosed for a privacy reason, and even these variables suffer typographical errors and inconsistent abbreviations [21, 59]. Correcting bias from imperfect record linkage requires a correctly specified model of linking errors [37, 39]. Our proposed method avoids these practical difficulties, and remains valid even when identification is possible.

2.3. *Hartley-type empirical process.* The empirical measure is a fundamental object in empirical process theory. This cannot be computed in our setting because of nonselected items and unidentified duplicated selection. As an alternative, we propose to study Hartley's estimator [25, 26] of a distribution function in place of the empirical measure.

Hartley's estimator [25, 26] was originally proposed for estimation of population total and average in multiple-frame surveys in sampling theory where multiple samples are drawn from overlapping sampling frames. Viewing sampling frames as data sources in our context, Hartley's estimator of the sample average $\mathbb{P}_N X$ of X when $J = 2$ is defined as

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

where the weight function ρ for duplicated selection is given by

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

for positive constants $c^{(1)}, c^{(2)}$ with $c^{(1)} + c^{(2)} = 1$. Duplicated selection and missing observations are properly addressed by the weight function $\rho(v)$ and the inverse probability weights, respectively. In fact, this estimator is unbiased for $E(X)$ because $\rho^{(1)}(v) + \rho^{(2)}(v) = 1$ for all v and $E[R_i^{(j)} | X_i, V_i, N^{(j)}, n^{(j)}] = \pi^{(j)}(V_i)$. Moreover, identification of duplicated items is not necessary to compute this estimator because the two sums

$$\mathbb{P}_N^{\mathrm{H}} X = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i^{(1)} \rho^{(1)}(V_i)}{\pi^{(1)}(V_i)} X_i + \frac{1}{N} \sum_{i=1}^{N} \frac{R_i^{(2)} \rho^{(2)}(V_i)}{\pi^{(2)}(V_i)} X_i, \tag{2.1}$$

in $\mathbb{P}_N^{\mathrm{H}} X$ can be computed separately based on each subsample.

Motivated by Hartley's estimator, we define the *Hartley-type empirical measure* (H-empirical measure) for $J = 2$ by

$$\mathbb{P}_N^{\mathrm{H}} \equiv \frac{1}{N} \sum_{i=1}^{N} \bigg(\frac{R_i^{(1)} \rho^{(1)}(V_i)}{\pi^{(1)}(V_i)} + \frac{R_i^{(2)} \rho^{(2)}(V_i)}{\pi^{(2)}(V_i)} \bigg) \delta_{(X_i, V_i)}.$$

This is an unbiased estimator of the empirical measure $\mathbb{P}_N \equiv N^{-1} \sum_{i=1}^{N} \delta_{(X_i, V_i)}$ given $(X_i, V_i), i = 1, \ldots, N$. Note, however, that $\mathbb{P}_N^{\mathrm{H}}$ is not a probability measure since point masses do not add up to 1 in general. The *Hartley-type empirical process* (H-empirical process) is defined by

$$\mathbb{G}_N^{\mathrm{H}} = \sqrt{N} \big(\mathbb{P}_N^{\mathrm{H}} - \tilde{P}_0 \big).$$

When there are more than two sources, we define the weight function $\rho = (\rho^{(1)}, \ldots, \rho^{(J)}) : \mathcal{V} \mapsto [0,1]^J$ that is constant on a mutually exclusive subset of $\mathcal{V}$ determined by $\mathcal{V}^{(j)}$'s:

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

with $j, k_1, \ldots, k_l$ all different and $\sum_{j=1}^{J} \rho^{(j)}(v) = 1$. The H-empirical measure is defined by

$$\mathbb{P}_N^{\mathrm{H}} \equiv \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \frac{R_i^{(j)} \rho^{(j)}(V_i)}{\pi^{(j)}(V_i)} \delta_{(X_i, V_i)}$$

and the H-empirical process is defined by $\mathbb{G}_N^{\mathrm{H}} = \sqrt{N}(\mathbb{P}_N^{\mathrm{H}} - \tilde{P}_0)$.

Let $\mathcal{F}$ be a class of measurable functions on $(\mathcal{X}, \mathcal{A})$ that serves as the index set for the H-empirical process. As a stochastic process indexed by $\mathcal{F}$, $\mathbb{G}_N^{\mathrm{H}}$ evaluated at $f \in \mathcal{F}$ is a random variable $\mathbb{G}_N^{\mathrm{H}} f = \sqrt{N}(\mathbb{P}_N^{\mathrm{H}} - \tilde{P}_0) f = \sqrt{N}(\mathbb{P}_N^{\mathrm{H}} f - \tilde{P}_0 f)$ where $\tilde{P}_0 f$ is the expectation of $f(X)$ under $\tilde{P}_0$, and $\mathbb{P}_N^{\mathrm{H}} f$ is the "expectation" of $f(X)$ under $\mathbb{P}_N^{\mathrm{H}}$ given by

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

We often omit variables of a function in "expectations" as in $\mathbb{P}_N^{\mathrm{H}} f$ and $\mathbb{G}_N^{\mathrm{H}} f$.

3. Limit theorems: Uniform WLLN and CLT. The U-LLN and U-CLT for the H-empirical process lay the groundwork for the analysis of merged data from multiple sources. The critical issue for establishing these theorems is multitiered dependence. This is not a difficult problem in the finite-population framework where only sampling indicators $R_i^{(j)}$ are random. For example, the two terms of $\mathbb{P}_N^{\mathrm{H}} X$ in (2.1) are independent in this framework, and each admits a finite-population CLT (e.g., [24]) to yield the sum of independent normal random variables as a limit [43]. A similar idea appears in the analysis of stratified samples. For the derivation of the U-CLT, [7] decomposed their weighted empirical process into stratum-wise empirical processes and showed their conditional weak convergence to independent Gaussian processes given data. Because strata do not overlap unlike our case, conditional independence automatically becomes unconditional to complete their proof. Unfortunately, this conditional argument is not valid in our setting due to dependence across overlapping data sources.

Our approach consists of two key ideas: (1) the decomposition of the H-empirical process into data sources with centering by appropriate variables, and (2) bootstrap asymptotics for establishing unconditional asymptotic normality. Our decomposition ensures unconditional independence, and bootstrap asymptotics bridges unconditional and conditional convergence.

Our decomposition emulates two stages of the sampling procedure:

$$\begin{aligned}\mathbb{G}_N^{\mathrm{H}} &= \sqrt{N}(\mathbb{P}_N^{\mathrm{H}} - \tilde{P}_0) \\ &= \sqrt{N}(\mathbb{P}_N - \tilde{P}_0) + \sqrt{N}(\mathbb{P}_N^{\mathrm{H}} - \mathbb{P}_N).\end{aligned}$$

The first term is the empirical process $\mathbb{G}_N = \sqrt{N}(\mathbb{P}_N - \tilde{P}_0)$ for the i.i.d. sample which corresponds to sampling from population at the first stage. This process weakly converges to the Brownian bridge by the U-CLT for the i.i.d. sample. The second term corresponding to sampling from data sources is further decomposed. Note that $\mathbb{P}_N f = \sum_{j=1}^{J} \mathbb{P}_N \rho^{(j)}(V) f(X)$ by the fact that $\sum_{j=1}^{J} \rho^{(j)}(v) = 1$ for every v. Combining this with the decomposition of $\mathbb{P}_N^{\mathrm{H}}$ in (2.1) with a general J

yields

$$(\mathbb{P}_N^{\mathrm{H}} - \mathbb{P}_N) f = \sum_{j=1}^{J} \frac{1}{N} \sum_{i=1}^{N} \Big(\frac{R_i^{(j)}}{\pi^{(j)}(V_i)} - 1 \Big) \rho^{(j)}(V_i) f(X_i)$$

$$\equiv \sum_{j=1}^{J} (\mathbb{P}_N^{\mathrm{H},(j)} - \mathbb{P}_N) \rho^{(j)} f.$$

As in the finite-population framework, the conditional covariance of $(\mathbb{P}_N^{\mathrm{H},(j)} - \mathbb{P}_N)\rho^{(j)} f$ with different j's is zero given data $(X_i, V_i), i = 1, \ldots, N$, because sampling from different data sources (i.e., $R^{(j)}$s and $R^{(j')}$s) is independent. Moreover, their conditional expectations given data are also zero because $E[R_i^{(j)} | X_i, V_i, N^{(j)}, n^{(j)}] = \pi^{(j)}(V_i)$. It follows from the total law of covariance i.e., $\mathrm{Cov}(X, Y) = E[\mathrm{Cov}(X, Y|Z)] + \mathrm{Cov}(E[X|Z], E[Y|Z])$ that any two of summands in the last display are uncorrelated. The same argument applies to the relationship between each summand and $\sqrt{N}(\mathbb{P}_N - \tilde{P}_0) f$. Hence we obtain the decomposition of $\mathbb{G}_N^{\mathrm{H}}$ into $J + 1$ uncorrelated pieces:

$$\mathbb{G}_N^{\mathrm{H}} f = \sqrt{N}(\mathbb{P}_N - \tilde{P}_0) f + \sum_{j=1}^{J} \sqrt{N} (\mathbb{P}_N^{\mathrm{H},(j)} - \mathbb{P}_N) \rho^{(j)} f.$$

If we show each summand converges to a Gaussian process, the limiting process of $\mathbb{G}_N^{\mathrm{H}}$ is the sum of $J + 1$ independent Gaussian processes.

To establish weak convergence of the second term in the last display, we adopt the bootstrap asymptotic theory. The key observation is to view sampling from a data source j as a single realization of the m-out-of-n bootstrap with $m = n^{(j)}$ and $n = N^{(j)}$ where a bootstrap sample of size m is drawn from a sample of size n without replacement. To see this, rewrite $(\mathbb{P}_N^{\mathrm{H},(j)} - \mathbb{P}_N)\rho^{(j)} f$ by $(N^{(j)}/N)(\hat{\mathbb{P}}_{n^{(j)}}^{(j)} - \mathbb{P}_{N^{(j)}}^{(j)})\rho^{(j)} f$ where

$$\hat{\mathbb{P}}_{n^{(j)}}^{(j)} \equiv \frac{1}{n^{(j)}} \sum_{i=1}^{N^{(j)}} R_{(j),i}^{(j)} \delta_{(X_{(j),i}, V_{(j),i})}, \qquad \mathbb{P}_{N^{(j)}}^{(j)} \equiv \frac{1}{N^{(j)}} \sum_{i=1}^{N^{(j)}} \delta_{(X_{(j),i}, V_{(j),i})}. \tag{3.1}$$

Here, we enumerate the items within data source j. Focusing on source j, $\mathbb{P}_{N^{(j)}}^{(j)} \rho^{(j)} f$ is the sample mean of $\rho^{(j)}(V) f(X)$ before sampling at the second stage while $\hat{\mathbb{P}}_{n^{(j)}}^{(j)} \rho^{(j)} f$ is the sample mean after sampling. In view of the m-out-of-n bootstrap, the former is an average in the original sample while the latter is a bootstrap average, and hence their difference is expected to yield asymptotic normality with appropriate scaling. Although $m/n = n^{(j)}/N^{(j)} \to p^{(j)} \neq 0$ unlike the usual m-out-of-n bootstrap method, asymptotics in our case can be treated as the special case of the exchangeably weighted bootstrap studied by Præstgaard and Wellner [46]. The theory of [46] emphasized conditional weak convergence, but it

is not difficult to extend their proof to unconditional one. Accordingly, we obtain the sum of independent Gaussian processes as the limit of $\mathbb{G}_N^{\mathrm{H}}$. In Section A of the Supplementary Material [51], we make this heuristic argument rigorous.

Below we write P^* and E^* to mean outer probability of P^∞ and expectation with respect to P^*. Since empirical process theory concerns the supremum of random elements, we use these notation to take care of measurability issues. For more details, see Section 1.2 of [58]. A reader not interested in technical details can replace these by $\tilde{P}_0$ and E without harm.

3.1. *Uniform law of large numbers.* The U-LLN holds for the empirical measure $\mathbb{P}_N$ in the i.i.d. setting if the index set $\mathcal{F}$ is a Glivenko–Cantelli class (see, e.g., page 81 of [58]). This Glivenko–Cantelli property is sufficient for the U-LLN for merged data from multiple sources. The following result is obtained by applying the bootstrap U-LLN [58] to our decomposition of $\mathbb{G}_N^{\mathrm{H}}$.

THEOREM 3.1. *Suppose that $\mathcal{F}$ is P_0-Glivenko–Cantelli. Then*

$$\|\mathbb{P}_N^{\mathrm{H}} - \tilde{P}_0\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |(\mathbb{P}_N^{\mathrm{H}} - \tilde{P}_0)f| \to_{P^*} 0,$$

where $\|\ell\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\ell(f)|$ for a functional ℓ on $\mathcal{F}$.

3.2. *Uniform central limit theorem.* The empirical process $\mathbb{G}_N$ in the i.i.d. setting weakly converges to a Gaussian process if the index set $\mathcal{F}$ is a Donsker class (see, e.g., page 81 of [58]). This Donsker property is sufficient for the U-CLT for the H-empirical process $\mathbb{G}_N^{\mathrm{H}}$. This is an expected consequence from bootstrap asymptotics which does not need additional conditions.

THEOREM 3.2. *Suppose that $\mathcal{F}$ is P_0-Donsker. Then*

$$\mathbb{G}_N^{\mathrm{H}}(\cdot) \rightsquigarrow \mathbb{G}^{\mathrm{H}}(\cdot) \equiv \mathbb{G}(\cdot) + \sum_{j=1}^{J} \sqrt{\nu^{(j)}} \sqrt{\frac{1-p^{(j)}}{p^{(j)}}} \mathbb{G}^{(j)}(\rho^{(j)}\cdot)$$

in the class $\ell^\infty(\mathcal{F})$ of uniformly bounded functionals on $\mathcal{F}$ where the P_0-Brownian bridge process $\mathbb{G}$ and the $P_0^{(j)}$-Brownian bridge processes $\mathbb{G}^{(j)}$ are independent. The covariance function $\upsilon(\cdot,\cdot) = \mathrm{Cov}(\mathbb{G}^{\mathrm{H}}\cdot, \mathbb{G}^{\mathrm{H}}\cdot)$ on $\mathcal{F}\times\mathcal{F}$ is

$$\upsilon(f,g) = \mathrm{Cov}_0(f,g) + \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \mathrm{Cov}_0^{(j)}(\rho^{(j)}f, \rho^{(j)}g),$$

where Cov_0 and $\mathrm{Cov}_0^{(j)}$ are covariances under P_0 and $P_0^{(j)}$, respectively.

The asymptotic variance here admits natural interpretations. Consider $\mathbb{G}_N^{\mathrm{H}} f$ for estimation of $P_0 f$ for instance. Its asymptotic variance is

$$\mathrm{AV}(\mathbb{G}_N^{\mathrm{H}} f) = \underbrace{\mathrm{Var}_0\{f(X)\}}_{\text{population variance}} + \sum_{j=1}^{J} \underbrace{\nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \mathrm{Var}_0^{(j)}\{\rho^{(j)}(V) f(X)\}}_{\text{design variance from source } j},$$

where $\mathrm{Var}_0(f) = \mathrm{Cov}_0(f, f)$ and $\mathrm{Var}_0^{(j)}(f) = \mathrm{Cov}_0^{(j)}(f, f)$. The first and second terms correspond to sampling from population and data sources respectively. If we would obtain the i.i.d. sample instead, the asymptotic variance is only the first term $\mathrm{Var}_0\{f(X)\}$. This can be obtained from our formula if we would sample all items from each data source (i.e., $p^{(j)} = 1$). This implies that as long as we sample all the items at the second stage, combining multiple datasets does not increase the difficulty of estimation. If data source j is large (i.e., $\tilde{P}_0(V \in \mathcal{V}^{(j)}) = \nu^{(j)}$ is large), its contribution to asymptotic variance becomes larger. Each quantity in the variance formula is easily estimated by Hartley's estimator of moments (see also Section G of the Supplementary Material [51] for variance estimators for several regression models).

REMARK 3.1. In Theorems 3.1 and 3.2, we assume Glivenko–Cantelli and Donsker properties of $\mathcal{F}$ with respect to P_0 in order to emphasize that these properties in the i.i.d. setting are sufficient for our setting. A brief inspection of our proof reveals that our theorems hold valid for $\tilde{P}_0$-Glivenko–Cantelli and $\tilde{P}_0$-Donsker classes of functions defined on $\mathcal{W} = \mathcal{X} \times \mathcal{U}$.

3.2.1. *Finite-population sampling*. Finite-population sampling concerns randomness only from the selection of units, and is often of interest in sampling theory. As expected from our interpretation of asymptotic variance in Theorem 3.2, we only obtain design variance from sources in this framework.

COROLLARY 3.1. *Suppose that $\mathcal{F}$ is P_0-Donsker. Then*

$$\mathbb{G}_N^{\mathrm{H,fin}}(\cdot) \equiv \sqrt{N}(\mathbb{P}_N^{\mathrm{H}} - \mathbb{P}_N)(\cdot) \rightsquigarrow \mathbb{G}^{\mathrm{H,fin}}(\cdot) \equiv \sum_{j=1}^{J} \sqrt{\nu^{(j)}} \sqrt{\frac{1-p^{(j)}}{p^{(j)}}} \mathbb{G}^{(j)}(\rho^{(j)} \cdot)$$

in $\ell^\infty(\mathcal{F})$ conditionally on $(X_1, V_1), (X_2, V_2) \ldots$, with the covariance function $\upsilon^{\mathrm{fin}}(\cdot, \cdot) = \mathrm{Cov}(\mathbb{G}^{\mathrm{H,fin}} \cdot, \mathbb{G}^{\mathrm{H,fin}} \cdot)$ on $\mathcal{F} \times \mathcal{F}$ given by

$$\upsilon^{\mathrm{fin}}(f, g) = \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \mathrm{Cov}_0^{(j)}(\rho^{(j)} f, \rho^{(j)} g).$$

3.2.2. *Bernoulli sampling*. Sampling without replacement is often replaced by Bernoulli sampling for mathematical convenience. To see its consequence, we consider Bernoulli sampling within sources where selections from source j are i.i.d. Bernoulli($p^{(j)}$). Data from the same source then become independent, but dependence remains between datasets from overlapping sources. We write $\mathbb{G}_N^{\mathrm{H,Ber}}$ for the H-empirical process in this case.

THEOREM 3.3. *Suppose that $\mathcal{F}$ is P_0-Donsker. Then $\mathbb{G}_N^{\mathrm{H,Ber}} \rightsquigarrow \mathbb{G}^{\mathrm{H,Ber}}$ in $\ell^\infty(\mathcal{F})$ where $\mathbb{G}^{\mathrm{H,Ber}}$ is the zero-mean Gaussian process $\mathbb{G}^{\mathrm{H,Ber}}$ with covariance function $\upsilon^{\mathrm{Ber}}(\cdot,\cdot) = \mathrm{Cov}(\mathbb{G}^{\mathrm{H,Ber}}\cdot, \mathbb{G}^{\mathrm{H,Ber}}\cdot)$ on $\mathcal{F}\times\mathcal{F}$ given by*

$$\upsilon^{\mathrm{Ber}}(f,g) = \mathrm{Cov}_0(f,g) + \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} P_0^{(j)}\{(\rho^{(j)})^2 fg\}.$$

Bernoulli sampling yields larger asymptotic variance than sampling without replacement. As expected from the decomposition of the asymptotic variance, the difference appears only in the design variances.

COROLLARY 3.2 (Finite-population correction). *The asymptotic variance is smaller when subsamples from sources are obtained from sampling without replacement than from Bernoulli sampling. In particular,*

$$\mathrm{AV}(\mathbb{G}_N^{\mathrm{H}} f) = \mathrm{AV}(\mathbb{G}_N^{\mathrm{H,Ber}} f) - \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \{P_0^{(j)} \rho^{(j)}(V) f(X)\}^2.$$

3.2.3. *Optimal ρ*. We derive the optimal weight function ρ based on our U-CLT. We propose the use of the optimal ρ under Bernoulli sampling which only involves $p^{(j)}$ determined by design. The optimal ρ under sampling without replacement involves an estimand itself and should differ from parameter to parameter. We show the optimal choice under Bernoulli sampling works well under sampling without replacement in simulation studies in Section 5.

PROPOSITION 3.1 (Optimal ρ under Bernoulli sampling). *Let $f : \mathcal{X} \to \mathbb{R}^k$ be arbitrary with $P_0 f^2 < \infty$. Let $od(p) = (1-p)/p$. When $J = 2$, the optimal function ρ that minimizes the asymptotic variance of $\mathbb{G}_N^{\mathrm{H,Ber}} f$ has*

$$c^{(1)} = \frac{od(p^{(2)})}{od(p^{(1)}) + od(p^{(2)})},$$

$$c^{(2)} = \frac{od(p^{(1)})}{od(p^{(1)}) + od(p^{(2)})}.$$

When $J \geq 2$, the optimal function ρ that minimizes the asymptotic variance of $\mathbb{G}_N^{\mathrm{H,Ber}} f$ has (1) $c^{(j)}_{k_1,\ldots,k_l} = 0$ *if* $p^{(j)} < 1$ *and* $p^{(k_m)} = 1$ *for some* m, (2) *arbitrary* $c^{(j)}_{k_1,\ldots,k_l}$ *if* $p^{(j)} = 1$ *and* (3)

$$c^{(j)}_{k_1,\ldots,k_l} = \frac{\prod_{m=1}^{l} od(p^{(k_m)})}{\prod_{m=1}^{l} od(p^{(k_m)}) + od(p^{(j)}) \sum_{n=1}^{l} \prod_{m=1}^{l} od(p^{(k_m)})/od(p^{(k_n)})},$$

if $p^{(j)}, p^{(k_m)} < 1$, $m = 1, \ldots, l$.

For sampling without replacement, we treat the case $J = 2$ only. The general case can be similarly derived via quadratic programming.

PROPOSITION 3.2 (Optimal ρ under sampling without replacement). *Let* $f : \mathcal{X} \to \mathbb{R}^k$ *be a function with* $P_0 f^2 < \infty$. *Let* $Y_f \equiv f(X) I\{V \in \mathcal{V}^{(1)} \cap \{\mathcal{V}^{(2)}\}^c\}$ *and* $Z_f \equiv f(X) I\{V \in \mathcal{V}^{(1)} \cap \mathcal{V}^{(2)}\}$. *Define*

$$c_f \equiv \frac{-\nu^{(1)} od(p^{(1)}) P_0^{(1)} Y_f P_0^{(1)} Z_f + \nu^{(2)} od(p^{(2)}) \{P_0^{(2)} Y_f P_0^{(2)} Z_f - \mathrm{Var}_0^{(2)}(Z_f)\}}{\nu^{(1)} od(p^{(1)}) \mathrm{Var}_0^{(1)}(Z_f) + \nu^{(2)} od(p^{(2)}) \mathrm{Var}_0^{(2)}(Z_f)}.$$

When $J = 2$, the optimal function ρ that minimizes the asymptotic variance of $\mathbb{G}_N^{\mathrm{H}} f$ has $c^{(1)} = 0 \vee c_f \wedge 1$ and $c^{(2)} = 1 - c$.

In a finite-population framework, [41] derived optimal ρ for general complex surveys. Their optimal ρ agrees with ours under Bernoulli sampling, but they differ under sampling without replacement. The difference is due to their probabilistic framework where [41] minimizes variance of $\mathbb{P}_N^{\mathrm{H}} f$ (which is zero in the limit) rather than the asymptotic variance of $\mathbb{G}_N^{\mathrm{H}} f$.

4. Calibration. The H-empirical process is computed from selected units only. If information on auxiliary variables V are available for non-selected units, calibration methods improve efficiency of our estimator. The key idea for calibration is that a statistic computed from sampled units (e.g., $\mathbb{P}_N^{\mathrm{H}} V$) is approximately equal to a statistic computed from all units (e.g., $\mathbb{P}_N V$). Adjusting weights in $\mathbb{P}_N^{\mathrm{H}}$ that induce similarity between two statistics makes selected units more representative of the population. Different methods use different pairs of two statistics. Below, we first introduce the extension of [47] to a general $J \geq 2$ and then propose our method.

The original calibration [16] ((2.3) of page 377) equates the Horvitz–Thompson estimator [29] of $\tilde{P}_0 V$ and sample average $\mathbb{P}_N V$ in order to improve the Horvitz–Thompson estimator of $P_0 X$. Along the same line, [47] imposed a constraint on Hartley's estimator $\mathbb{P}_N^{\mathrm{H}} V$ and sample average $\mathbb{P}_N V$ to improve $\mathbb{P}_N^{\mathrm{H}} X$ when $J = 2$. For a general J, we consider as its extension the following *calibration equation*:

$$\mathbb{P}_N^{\mathrm{H}} G(V^T \alpha) V = \mathbb{P}_N V, \tag{4.1}$$

with a solution $\hat{\alpha}_N^{\mathrm{c}}$. Here, G is a fixed function (see [16] for some choice of G). Using $G(V^T\hat{\alpha}_N^{\mathrm{c}})$, the calibrated H-empirical measure is defined as

$$\mathbb{P}_N^{\mathrm{H,c}}(\cdot) \equiv \mathbb{P}_N^{\mathrm{H}} G\big(V^T\hat{\alpha}_N^{\mathrm{c}}\big)(\cdot) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)} G\big(V_i^T\hat{\alpha}_N^{\mathrm{c}}\big)\delta_{(X_i,V_i)},$$

and the calibrated H-empirical process is defined as $\mathbb{G}_N^{\mathrm{H,c}} \equiv \sqrt{N}(\mathbb{P}_N^{\mathrm{H,c}} - \tilde{P}_0)$. Other variants in [47] can be extended by changing the range of summation. For example, if we replace V by a vector with elements $VI_{\mathcal{V}^{(j)}}(V)$, $j = 1, \ldots, J$, in (4.1), we obtain *data-source-specific calibration*

$$\frac{1}{N}\sum_{i:V_i\in\mathcal{V}^{(j)}}\sum_{j=1}^{J}\frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)} G\big(V_i^T\alpha^{(j)}\big)V_i = \frac{1}{N}\sum_{i:V_i\in\mathcal{V}^{(j)}} V_i, \qquad j = 1, \ldots, J.$$

The left-hand side is computed from all selected units that belong to source j.

Our proposed method exploits the asymptotic variance formula $\upsilon(f, f)$ in Theorem 3.2. We target the reduction of design variances in

$$\mathrm{Var}_0^{(j)}\big\{\rho^{(j)}(V)f(X)\big\} = P_0^{(j)}\big\{\rho^{(j)}(V)f(X) - P_0^{(j)}\rho^{(j)}(V)f(X)\big\}^{\otimes 2}.$$

The key observations are (1) the conditional variance is obtained from the sample from the same source (units with $R^{(j)} = 1$), and (2) variables of interest are $\rho^{(j)}(V)f(X) - P_0^{(j)}\rho^{(j)}(V)f(X)$. Our method is thus characterized by the following three points: (1) calibration is carried out within a subsample from the same source, (2) variables used are $\rho^{(j)}(V)V$ with centering and (3) Horvitz–Thompson estimators are equated with sample averages. To be specific, we propose the *sample-specific calibration equation*

$$\frac{1}{N^{(j)}}\sum_{i:V_i\in\mathcal{V}^{(j)}}\frac{R_i^{(j)}G_{\alpha^{(j)}}^{(j)}(V_i)}{\pi^{(j)}(V_i)}\big\{\rho^{(j)}(V_i)V_i - \mathbb{P}_{N^{(j)}}^{(j)}\rho^{(j)}(V)V\big\} = 0, \tag{4.2}$$

$j = 1, \ldots, J$ with solution $\hat{\alpha}_N^{\mathrm{sc}} = (\hat{\alpha}_N^{\mathrm{sc},(1)}, \ldots, \hat{\alpha}_N^{\mathrm{sc},(J)})^T$ where

$$G_\alpha^{(j)}(v) \equiv G\big[\big\{\rho^{(j)}(v)v - \mathbb{P}_{N^{(j)}}^{(j)}\rho^{(j)}(V)V\big\}^T\alpha^{(j)}\big].$$

The right-hand side of (4.2) is the average of empirically centered variables, and hence equals zero. The left-hand side is computed from selected items from source j in contrast to data-source-specific calibration that uses all items sampled from both source j and its overlapping sources. We define the *H-empirical measure with sample-specific calibration* by

$$\mathbb{P}_N^{\mathrm{H,sc}} \equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)} G_{\hat{\alpha}_N^{\mathrm{sc},(j)}}^{(j)}(V_i)\delta_{(X_i,V_i)},$$

and the corresponding H-empirical process by $\mathbb{G}_N^{\mathrm{H,sc}} \equiv \sqrt{N}(\mathbb{P}_N^{\mathrm{H,sc}} - \tilde{P}_0)$.

We assume the following condition for calibration methods.

CONDITION 4.1. (a) $\hat{\alpha}_N^{\mathrm{c}}$ and $\hat{\alpha}_N^{\mathrm{sc}}$ are solutions of (4.1) and (4.2).

(b) $V \in \mathbb{R}^k$ has bounded support with $\mathcal{V}^{(j)} \neq \{0\}$, $j = 1, \ldots, J$.

(c) G is a strictly increasing, continuously differentiable, bounded function on $\mathbb{R}$ such that $G(0) = 1$. Its derivative $\dot{G}$ is strictly positive and bounded.

(d) $\tilde{P}_0 V^{\otimes 2}$ and every $\mathrm{Var}_0^{(j)}\{\rho^{(j)}(V)V\}$ satisfying $P_0^{(j)}\rho^{(j)}(V) > 0$ are finite and positive definite.

Condition 4.1(a) ensures the existence of solutions to calibration equations. Under conditions (b)–(d), probability of their existence with the choice $G(x) = 1 + x$ tends to 1 as $N \to \infty$. When V is bounded, $G(x) = 1 + x$ can be considered as a bounded function that satisfies (c). In this case,

$$\hat{\alpha}_N^{\mathrm{sc}(j)} = \Biggl\{ \frac{1}{N^{(j)}} \sum_{i: V_i \in \mathcal{V}^{(j)}} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \bigl(V_i^{\rho^{(j)}}\bigr)^{\otimes 2} \Biggr\}^{-1} \frac{1}{N^{(j)}} \sum_{i: V_i \in \mathcal{V}^{(j)}} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} V_i^{\rho^{(j)}},$$

where $V_i^{\rho^{(j)}} \equiv \rho^{(j)}(V_i)V_i - \mathbb{P}_{N^{(j)}}^{(j)}\rho^{(j)}(V)V$. The probability of the existence of the matrix inverse above tends to 1 due to (d). A similar argument applies to $\hat{\alpha}_N^{\mathrm{c}}$. Note that the choice of G does not affect the limiting processes in the uniform CLT for calibrated H-empirical processes below.

THEOREM 4.1. *Suppose $\mathcal{F}$ is P_0-Donsker with $\|P_0\|_{\mathcal{F}} < \infty$. Under Condition 4.1,*

$$\mathbb{G}_N^{\mathrm{H,c}}(\cdot) \rightsquigarrow \mathbb{G}^{\mathrm{H,c}}(\cdot) \equiv \mathbb{G}(\cdot) + \sum_{j=1}^{J} \sqrt{\nu^{(j)}} \sqrt{\frac{1 - p^{(j)}}{p^{(j)}}} \mathbb{G}^{(j)}\bigl(\rho^{(j)} I - Q_{\mathrm{c}}^{(j)}\bigr)(\cdot),$$

$$\mathbb{G}_N^{\mathrm{H,sc}}(\cdot) \rightsquigarrow \mathbb{G}^{\mathrm{H,sc}}(\cdot) \equiv \mathbb{G}(\cdot) + \sum_{j=1}^{J} \sqrt{\nu^{(j)}} \sqrt{\frac{1 - p^{(j)}}{p^{(j)}}} \mathbb{G}^{(j)}\bigl(\rho^{(j)} I - Q_{\mathrm{sc}}^{(j)}\bigr)(\cdot),$$

in $\ell^\infty(\mathcal{F})$. Here, $\mathbb{G}$ and $\mathbb{G}^{(j)}$ are the same as in Theorem 3.2, I is the identity map, and $Q_{\mathrm{c}}^{(j)}$ and $Q_{\mathrm{sc}}^{(j)}$ are maps from the class of functions on $\mathcal{X}$ to the class of linear maps on $\mathcal{V}$ defined by

$$Q_{\mathrm{c}}^{(j)}(f)[v] = \tilde{P}_0\bigl(f(X)V^T\bigr)\bigl\{\tilde{P}_0 V^{\otimes 2}\bigr\}^{-1}\rho^{(j)}(v)v,$$

$$\begin{aligned} Q_{\mathrm{sc}}^{(j)}(f)[v] = {} & P_0^{(j)}\bigl\{\rho^{(j)}(V)f(X)\bigl(\rho^{(j)}(V)V - P_0^{(j)}\rho^{(j)}(V)V\bigr)^T\bigr\} \\ & \times \bigl\{\mathrm{Var}_0^{(j)}\bigl(\rho^{(j)}(V)V\bigr)\bigr\}^{-1}\bigl\{\rho^{(j)}(v)v - P_0^{(j)}\rho^{(j)}(V)V\bigr\} I\bigl\{v \in \mathcal{V}^{(j)}\bigr\}. \end{aligned}$$

Covariance functions $\upsilon^{\#}(\cdot,\cdot) = \mathrm{Cov}(\mathbb{G}^{\mathrm{H},\#}\cdot, \mathbb{G}^{\mathrm{H},\#}\cdot)$ on $\mathcal{F} \times \mathcal{F}$, $\# \in \{\mathrm{c},\mathrm{sc}\}$ are

$$\begin{aligned} \upsilon^{\#}(f,g) = {} & \mathrm{Cov}_0(f,g) \\ & + \sum_{j=1}^{J} \nu^{(j)} \frac{1 - p^{(j)}}{p^{(j)}} \mathrm{Cov}_0^{(j)}\bigl(\rho^{(j)} f - Q_{\#}^{(j)}(f), \rho^{(j)} g - Q_{\#}^{(j)}(g)\bigr). \end{aligned}$$

To compare above methods, define the class $\mathcal{C}$ of estimators of $P_0 f$ for arbitrary f with $P_0 f^{\otimes 2} < \infty$ whose asymptotic variance takes the form of

$$\mathrm{Var}_0\{f(X)\} + \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \mathrm{Var}_0^{(j)}[\rho^{(j)}(V) f(X) - L_f^{(j)}\{\rho^{(j)}(V)V\}],$$

where $L_f^{(j)}(v)$ is a linear function of v that depends on f. Note that calibration and the sample-specific calibration have $L_f^{(j)}\{\rho^{(j)}(v)v\} = Q_{\mathrm{c}}^{(j)}(f)[v]$ and $L_f^{(j)}\{\rho^{(j)}(v)v\} = Q_{\mathrm{sc}}^{(j)}(f)[v]$. The optimal $L_f^{(j)}\{\rho^{(j)}(v)v\}$ is the orthogonal projection of $\rho^{(j)}(v)f(x)$ onto the linear span of $\rho^{(j)}(v)v - P_0^{(j)}\{\rho^{(j)}(V)V\}$ with respect to the pseudo-metric $d^{(j)}(f,g) = \{\mathrm{Var}_0^{(j)}(f-g)\}^{1/2}$. This is exactly $L_f^{(j)}\{\rho^{(j)}(v)v\} = Q_{\mathrm{sc}}^{(j)}(f)[v]$. Thus, we obtain the following theorem.

THEOREM 4.2. *Sample-specific calibration is optimal among $\mathcal{C}$ with improved asymptotic variance over a noncalibrated estimator*:

$$\mathrm{AV}\big(\mathbb{G}_N^{\mathrm{H,sc}} f\big) = \mathrm{AV}\big(\mathbb{G}_N^{\mathrm{H}} f\big) - \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \mathrm{Var}_0^{(j)}\big(Q_{\mathrm{sc}}^{(j)} f\big).$$

The performance of methods based on [47] depends on specific situations. See our simulation in Section 6.

5. Applications to infinite-dimensional *M*-estimation. An estimator in a statistical model is often characterized as a maximizer of a criterion function or a zero of estimating equations. The former estimator is called an M-estimator and the latter a Z-estimator. A canonical example for both cases is the maximum likelihood estimator (MLE) which maximizes likelihood and solves likelihood equations. In the i.i.d. setting, empirical process theory plays a major role in studying both estimators in a general setting where parameters are infinite-dimensional [18, 44, 56, 58]. In this section, we apply H-empirical process results to study limiting properties of infinite-dimensional M- and Z-estimation for data integration.

Suppose $\mathcal{P}$ is the collection of probability measures P_θ on $(\mathcal{X}, \mathcal{A})$ parametrized by $\theta \in \Theta$ where Θ is a subset of a Banach space $(\mathcal{B}, \|\cdot\|)$. The true distribution is $P_0 = P_{\theta_0} \in \mathcal{P}$. Let $\mathcal{M} = \{m_\theta : \theta \in \Theta\}$ be a set of criterion functions on $\mathcal{X}$. In the i.i.d. setting, the M-estimator is defined as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \mathbb{P}_N m_\theta(X).$$

Our proposed M-estimator $\hat{\theta}_N$ replaces the empirical measure by the H-empirical measure:

$$\hat{\theta}_N = \arg\max_{\theta\in\Theta} \mathbb{P}_N^{\mathrm{H}} m_\theta(X).$$

In the following, we establish consistency and rates of convergence of our M-estimator, while we consider Z-estimation for asymptotic normality. Treating two estimators interchangeably can be justified because the M-estimator often (nearly) solves estimating equations obtained from the criterion function. This relationship must be verified in each specific model.

5.1. *Consistency.* The following theorem concerns consistency of our proposed M-estimator. The key assumption is the Glivenko–Cantelli property of $\mathcal{M}$ by which our U-LLN applies.

THEOREM 5.1. *Suppose that $\mathcal{M}$ is P_0-Glivenko–Cantelli, and that for every $\epsilon > 0$, $P_0 m_{\theta_0} > \sup_{\theta:\|\theta-\theta_0\|>\epsilon} P_0 m_\theta$. Then*

$$\|\hat{\theta}_N - \theta_0\| \to_{P^*} 0.$$

In certain semiparametric models MLEs do not exist and nonparametric MLEs are considered as alternatives. In this case, the parameter space for optimization may not be the same as the original space, and consistency must be carefully proved based on properties of a specific model. Our U-LLN continues to be helpful for this purpose (see Example 5.3).

5.2. *Rate of convergence.* A rate of convergence appears in Condition 5.4 for one of our Z-theorems, namely, Theorem 5.4. In the i.i.d. case, convergence rates are often obtained by the peeling device ([1], see also Theorem 3.2.5 of [58]) together with maximal inequalities for the empirical process. Instead of obtaining maximal inequalities of H-empirical processes for different $\mathcal{M}$ each time, we directly compare maximal inequalities for the empirical and H-empirical processes to obtain the following theorem. This theorem ensures the same rate of convergence both in the i.i.d. setting and our setting. Below, we denote $a \lesssim b$ to mean $a \leq Kb$ for some constant $K \in (0, \infty)$.

THEOREM 5.2. *Suppose for every θ in a neighborhood of θ_0,*

$$P_0(m_\theta - m_{\theta_0}) \lesssim -\|\theta - \theta_0\|^2. \tag{5.1}$$

For every N and sufficiently small $\delta > 0$, it holds that

$$E^* \sup_{\|\theta-\theta_0\|<\delta} \left|\mathbb{G}_N(m_\theta - m_{\theta_0})\right| \lesssim \phi_N(\delta) \tag{5.2}$$

for functions ϕ_N such that $\delta \mapsto \phi_N(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ (not depending on N). If $\hat{\theta}_N \to_{P^*} \theta_0$ and $\mathbb{P}_N^{\mathrm{H}} m_{\hat{\theta}_N} \geq \mathbb{P}_N^{\mathrm{H}} m_{\theta_0} - O_{P^*}(r_N^{-2})$, then $r_N\|\hat{\theta}_N - \theta_0\| = O_{P^*}(1)$ for every r_N such that $r_N^2 \phi_N(1/r_N) \leq \sqrt{N}$ for every N.

5.3. *Infinite-dimensional Z-theorem.* We consider asymptotic distributions of our Z-estimators by extending two infinite-dimensional Z-theorems (Theorem 3.3.1 of [58] and Theorem 6.1 of [31]) in the i.i.d. setting to our setting. The first theorem concerns estimators with regular parametric rate of convergence. The second theorem specializes in semiparametric models with nonregular rate of convergence for nuisance parameters. The estimators are obtained by replacing $\mathbb{P}_N$ by $\mathbb{P}_N^{\mathrm{H}}$ in estimating equations. We also consider calibration methods in the previous section in these theorems.

5.3.1. *Parametric rates of convergence for nuisance parameters.* Let $\hat{\theta}_N$ and $\hat{\theta}_{N,\#}$ be estimators of θ obtained as solutions to the estimating equations given by

$$\begin{aligned}\bigl\|\Psi_N^{\mathrm{H}}(\theta)\bigr\|_{\mathcal{H}} &\equiv \bigl\|\mathbb{P}_N^{\mathrm{H}} B_\theta\bigr\|_{\mathcal{H}} = o_{P^*}\bigl(N^{-1/2}\bigr),\\ \bigl\|\Psi_{N,\#}^{\mathrm{H}}(\theta)\bigr\|_{\mathcal{H}} &\equiv \bigl\|\mathbb{P}_N^{\mathrm{H},\#} B_\theta\bigr\|_{\mathcal{H}} = o_{P^*}\bigl(N^{-1/2}\bigr), \qquad \# \in \{\mathrm{c},\mathrm{sc}\},\end{aligned}$$

respectively where B_θ is a map from some set $\mathcal{H}$ to $L_2(P_\theta)$ indexed by θ. Recall, for example, $\|\mathbb{P}_N^{\mathrm{H}} B_\theta\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}} |\mathbb{P}_N^{\mathrm{H}} B_\theta(h)|$ (see also Example 5.1). Let $\Psi(\theta) \equiv P_0 B_\theta$ and $\Psi_N(\theta) \equiv \mathbb{P}_N B_\theta$ be maps from Θ to $\ell^\infty(\mathcal{H})$. We assume the following.

CONDITION 5.1. For the true parameter $\theta_0 \in \Theta$, $\Psi(\theta_0) = 0$. The set $\{B_{\theta_0}(h) : h \in \mathcal{H}\}$ is P_0-Donsker and $\{(B_\theta - B_{\theta_0})(h) : \theta \in \Theta, h \in \mathcal{H}\}$ is P_0-Glivenko–Cantelli with an integrable envelope.

CONDITION 5.2. Suppose that Ψ is Fréchet differentiable at θ_0:

$$\bigl\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_0(\theta - \theta_0)\bigr\|_{\mathcal{H}} = o\bigl(\|\theta - \theta_0\|\bigr).$$

Moreover, $\dot{\Psi}_0$ is continuously invertible at θ_0 with inverse denoted as $\dot{\Psi}_0^{-1}$.

CONDITION 5.3. For any $\delta_N \downarrow 0$, the following stochastic equicontinuity condition holds at θ_0:

$$\sup_{\|\theta-\theta_0\|\le\delta_N} \left\| \frac{\sqrt{N}(\Psi_N - \Psi)(\theta) - \sqrt{N}(\Psi_N - \Psi)(\theta_0)}{1 + \sqrt{N}\|\theta - \theta_0\|} \right\|_{\mathcal{H}} = o_{P^*}(1).$$

Now we present the following infinite-dimensional Z-theorem.

THEOREM 5.3. *Suppose that Conditions 5.1–5.3 hold and that estimators* $\hat{\theta}_N, \hat{\theta}_{N,\#}$, *with* $\# \in \{\mathrm{c},\mathrm{sc}\}$ *are consistent for* θ_0. *Then*

$$\begin{aligned}\sqrt{N}(\hat{\theta}_N - \theta_0) &\rightsquigarrow -\dot{\Psi}_0^{-1}\mathbb{G}^{\mathrm{H}} B_{\theta_0},\\ \sqrt{N}(\hat{\theta}_{N,\#} - \theta_0) &\rightsquigarrow -\dot{\Psi}_0^{-1}\mathbb{G}^{\mathrm{H},\#} B_{\theta_0}.\end{aligned}$$

5.3.2. *Nonregular rates of convergence for nuisance parameters.* We focus on a semiparametric model $\mathcal{P} = \{p_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^p, \eta \in H\}$, the collection of densities on $(\mathcal{X}, \mathcal{A})$ where $\Theta \subset \mathbb{R}^p$, and H is a subset of a Banach space $(\mathcal{B}, \|\cdot\|)$. The true distribution is $P_0 = P_{\theta_0,\eta_0} \in \mathcal{P}$. Estimator $(\hat{\theta}_N, \hat{\eta}_N)$ solves the Hartley-type likelihood equations

$$\begin{aligned} \Psi^{\mathrm{H}}_{N,1}(\theta,\eta,\alpha) &= \mathbb{P}^{\mathrm{H}}_N \dot{\ell}_{\theta,\eta} = o_{P^*}\big(N^{-1/2}\big), \\ \Psi^{\mathrm{H}}_{N,2}(\theta,\eta,\alpha)[\underline{h}_0] &= \mathbb{P}^{\mathrm{H}}_N B_{\theta,\eta}[\underline{h}_0] = o_{P^*}\big(N^{-1/2}\big). \end{aligned} \tag{5.3}$$

Here, $\dot{\ell}_{\theta,\eta} \in \mathcal{L}^0_2(P_{\theta,\eta})^p$ is the score function for θ, and the score operator $B_{\theta,\eta} : \mathcal{H} \mapsto \mathcal{L}^0_2(P_{\theta,\eta})$ is the bounded linear operator mapping a direction h in some Hilbert space $\mathcal{H}$ of one-dimensional submodels for η along which $\eta \to \eta_0$ (see, e.g., [57] for review of semiparametric models). We write $B_{\theta,\eta}[\underline{h}] = (B_{\theta,\eta}(h_1), \ldots, B_{\theta,\eta}(h_p))^T$ for $\underline{h} = (h_1, \ldots, h_p)^T \in \mathcal{H}^p$, and $\underline{h}_0$ is defined in Condition 5.5 below. We also write $\dot{\ell}_0 = \dot{\ell}_{\theta_0,\eta_0}$ and $B_0 = B_{\theta_0,\eta_0}$. We assume the following.

CONDITION 5.4. An estimator $(\hat{\theta}_N, \hat{\eta}_N)$ of (θ_0, η_0) satisfies $|\hat{\theta}_N - \theta_0| = o_{P^*}(1)$, and $\|\hat{\eta}_N - \eta_0\| = O_{P^*}(N^{-\beta})$ for some $\beta > 0$, and solves the estimating equations (5.3) where $\mathbb{P}^{\mathrm{H}}_N$ may be replaced by $\mathbb{P}^{\mathrm{H},\#}_N$ with the corresponding estimators $(\hat{\theta}_{N,\#}, \hat{\eta}_{N,\#})$ where $\# \in \{\mathrm{c},\mathrm{sc}\}$.

CONDITION 5.5. There is an $\underline{h}_0 = (h_{0,1}, \ldots, h_{0,p})^T \in \mathcal{H}^p$ such that

$$P_0\big\{\big(\dot{\ell}_0 - B_0[\underline{h}_0]\big)B_0(h)\big\} = 0, \qquad \text{for all } h \in \mathcal{H}.$$

Furthermore, $I_0 \equiv P_0(\dot{\ell}_0 - B_0[\underline{h}_0])^{\otimes 2}$ is finite and nonsingular.

CONDITION 5.6. (1) For any $\delta_N \downarrow 0$ and $C > 0$,

$$\sup_{|\theta-\theta_0|\le\delta_N, \|\eta-\eta_0\|\le CN^{-\beta}} \big|\mathbb{G}_N(\dot{\ell}_{\theta,\eta} - \dot{\ell}_0)\big| = o_{P^*}(1),$$

$$\sup_{|\theta-\theta_0|\le\delta_N, \|\eta-\eta_0\|\le CN^{-\beta}} \big|\mathbb{G}_N(B_{\theta,\eta} - B_0)[\underline{h}_0]\big| = o_{P^*}(1).$$

(2) For some $\delta > 0$ classes $\{\dot{\ell}_{\theta,\eta} : |\theta-\theta_0| + \|\eta-\eta_0\| \le \delta\}$ and $\{B_{\theta,\eta}[\underline{h}_0] : |\theta-\theta_0| + \|\eta-\eta_0\| \le \delta\}$ are P_0-Glivenko–Cantelli and have integrable envelopes. Moreover, $\dot{\ell}_{\theta,\eta}$ and $B_{\theta,\eta}[\underline{h}_0]$ are continuous with respect to (θ,η) in $L_1(P_0)$.

CONDITION 5.7. For some $\alpha > 1$ satisfying $\alpha\beta > 1/2$ and for (θ,η) in the neighborhood $\{(\theta,\eta) : |\theta-\theta_0| \le \delta_N, \|\eta-\eta_0\| \le CN^{-\beta}\}$,

$$\begin{aligned} &\big|P_0\big[\dot{\ell}_{\theta,\eta} - \dot{\ell}_0 + \dot{\ell}_0\big\{\dot{\ell}_0^T(\theta-\theta_0) + B_0(\eta-\eta_0)\big\}\big]\big| \\ &\quad = o\big(|\theta-\theta_0|\big) + O\big(\|\eta-\eta_0\|^\alpha\big), \\ &\big|P_0\big[(B_{\theta,\eta} - B_0)[\underline{h}_0] + B_0[\underline{h}_0]\big\{\dot{\ell}_0^T(\theta-\theta_0) + B_0(\eta-\eta_0)\big\}\big]\big| \\ &\quad = o\big(|\theta-\theta_0|\big) + O\big(\|\eta-\eta_0\|^\alpha\big). \end{aligned}$$

We then obtain the following infinite-dimensional Z-theorem.

THEOREM 5.4. *Under Conditions* 4.1, 5.4–5.7,

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \rightsquigarrow \mathbb{G}^{\mathrm{H}}\tilde{\ell}_0 \sim N_p\big(0, \upsilon(\tilde{\ell}_0, \tilde{\ell}_0)\big),$$
$$\sqrt{N}(\hat{\theta}_{N,\#} - \theta_0) \rightsquigarrow \mathbb{G}^{\mathrm{H},\#}\tilde{\ell}_0 \sim N_p\big(0, \upsilon^{\#}(\tilde{\ell}_0, \tilde{\ell}_0)\big),$$

where $\# \in \{\mathrm{c},\mathrm{sc}\}$, *and* υ *and* $\upsilon^{\#}$ *are as defined in Theorems* 3.2 *and* 4.1.

5.4. *Examples*.

EXAMPLE 5.1 (Parametric model). Consider the parametric model $\{dP_\theta/d\mu = p_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ with a dominating measure μ. A natural estimator $\hat{\theta}_N$ of θ is a solution to the Hartley-type likelihood equation given by

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)}\dot{\ell}_\theta(X_i) = 0,$$

where $\dot{\ell}_\theta = d\log p_\theta/d\theta$. Let $\dot{\ell}_\theta(x) = (\dot{\ell}_{\theta,1}(x), \ldots, \dot{\ell}_{\theta,p}(x))^T$. For $\mathcal{H} = \{h_1, \ldots, h_p\}$, define the map by $h_i \mapsto B_\theta(h_i) = \dot{\ell}_{\theta,i}(x)$. Then the above estimating equation can be written as $\|\Psi_N^{\mathrm{H}}(\theta)\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}}|\mathbb{P}_N^{\mathrm{H}}B_\theta(h)| = 0$. Square integrability of $\dot{\ell}_\theta$ for each θ under P_0 implies Condition 5.1. When the Fisher information matrix $I_0 \equiv P_0(\dot{\ell}_0^{\otimes 2})$ is invertible, and $\log p_\theta$ is twice differentiable with respect to θ in a neighborhood of θ_0, Condition 5.2 is satisfied. If we further assume $\log p_\theta$ is twice continuously differentiable in a neighborhood of θ_0 and Θ is compact, Condition 5.3 is met. Consistency follows from Theorem 5.1 if $\{B_\theta(h_i), i = 1, \ldots, p, \theta \in \Theta\}$ has an integrable envelope. Hence our first Z-theorem (Theorem 5.3) yields

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to_d \mathbb{G}^{\mathrm{H}}\tilde{\ell}_0 \sim N\left(0, I_0^{-1} + \sum_{j=1}^{J}\nu^{(j)}\frac{1-p^{(j)}}{p^{(j)}}\mathrm{Var}_0^{(j)}(\rho^{(j)}\tilde{\ell}_0)\right),$$

where $\tilde{\ell}_0 = I_0^{-1}\dot{\ell}_0$. The cases for calibration are similar.

EXAMPLE 5.2 (Regular semiparametric model with η as measure). Consider the semiparametric model $\mathcal{P} = \{p_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^p, \eta \in H\}$ where the nuisance parameter η is a measure. Several Z-theorems of the form of Theorem 5.3 were applied to this case [58] (see Section 3.3 of [58] in the i.i.d. setting and [7, 8, 52] for stratified samples). We obtain a similar result from Theorem 5.3 by following arguments in [52]. The score operator in this model is $B_{\theta,\eta} : L_2(\eta) \mapsto L_2(P_{\theta,\eta})$ and its adjoint operator is denoted as $B_{\theta,\eta}^* : L_2(P_{\theta,\eta}) \mapsto L_2(\eta)$. As in [58], we

assume $B_0^* B_0$ is continuously invertible and that Ψ has continuously invertible Fréchet derivative $\dot{\Psi}_0$ at (θ_0, η_0) with respect to (θ, η) of the form

$$\dot{\Psi}_{11}(\theta - \theta_0) = -P_0 \dot{\ell}_0 \dot{\ell}_0^T (\theta - \theta_0),$$

$$\dot{\Psi}_{12}(\eta - \eta_0) = -\int B_0^* \dot{\ell}_0 d(\eta - \eta_0),$$

$$\dot{\Psi}_{21}(\theta - \theta_0)h = -P_0 B_0 h \dot{\ell}_0^T (\theta - \theta_0), \qquad h \in L_2(\eta),$$

$$\dot{\Psi}_{22}(\eta - \eta_0)h = -\int B_0^* B_0 h d(\eta - \eta_0), \qquad h \in L_2(\eta).$$

Further assuming consistency and asymptotic equicontinuity (see [7, 8, 52, 58] for more details), Theorem 5.3 yields

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to_d \mathbb{G}^{\mathrm{H}} \tilde{\ell}_0 \sim N\left(0, I_0^{-1} + \sum_{j=1}^{J} \nu^{(j)} \frac{1 - p^{(j)}}{p^{(j)}} \mathrm{Var}_0^{(j)}\left(\rho^{(j)} \tilde{\ell}_0\right)\right),$$

where $\tilde{I}_0 = P_0[(I - B_0(B_0^* B_0)^{-1} B_0^*)\dot{\ell}_0 \dot{\ell}_0^T]$ is the efficient information for θ and $\tilde{\ell}_0 = \tilde{I}_0^{-1}(I - B_0(B_0^* B_0)^{-1} B_0^*)\dot{\ell}_0$ is the efficient influence function for θ in the i.i.d. setting.

EXAMPLE 5.3 (Cox model with right-censored data). Let $T \sim F$ be a failure time, and $Z = (Z_1, Z_2)$ be covariates. The Cox model specifies the relationship between covariates and the cumulative hazard function by

$$\Lambda(t|z) = \exp\left(\theta^T z\right)\Lambda(t),$$

where $\theta \in \mathbb{R}^p$ is the regression parameter, and Λ is the baseline cumulative hazard function. Under right censoring, we do not always observe T but observe $Y \equiv \min\{T, C\}$ and $\Delta \equiv I(T \leq C)$ where C is censoring time. We assume there is some constant τ such that $P(T > \tau) > 0$ and $P(C \geq \tau) = P(C = \tau) > 0$ (see [55] for other conditions). We assume sources are formed based on $V = (Y, \Delta, Z_2)$ and Z_1 is collected later. In the i.i.d. setting, a nonparametric likelihood for one observation is $\ell_{\theta,\Lambda}(y, \delta, z) = \log\{(e^{\theta^T z}\Lambda\{y\})^{\delta} e^{-\Lambda(y)e^{\theta^T z}}\}$ where $\Lambda\{t\}$ is the jump of Λ at t. The score for θ and the score operator $B_{\theta,\Lambda} : \mathcal{H} \mapsto L_2(P_{\theta,\Lambda})$ are

$$\dot{\ell}_{\theta,\Lambda}(y, \delta, z) = z\left\{\delta - e^{\theta^T z}\Lambda(y)\right\},$$

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

where $\mathcal{H}$ is the unit ball in the space $BV[0, \tau]$. Here, the score operator is obtained by differentiating ℓ_{θ,Λ_t} with respect to t at $t = 0$ where $d\Lambda_t = (1 + th)\, d\Lambda$. Our proposed estimator $(\hat{\theta}_N, \hat{\Lambda}_N)$ is the solution to $\mathbb{P}_N^{\mathrm{H}} \dot{\ell}_{\theta,\Lambda} = 0$ and $\mathbb{P}_N^{\mathrm{H}} B_{\theta,\Lambda}(h) = 0$, whereby $\hat{\theta}_N$ is the solution to the weighted partial likelihood equation and $\hat{\Lambda}_N$ is the weighted Breslow estimator (see, e.g., [7]). Consistency and conditions for asymptotic normality can be verified along the same line as in [52] by replacing

their weighted empirical process results by our H-empirical process results. Then Example 5.2 yields

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to_d \mathbb{G}^{\mathrm{H}}\tilde{\ell}_0 \sim N\left(0, I_0^{-1} + \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \mathrm{Var}_0^{(j)}\left(\rho^{(j)}\tilde{\ell}_0\right)\right).$$

Here the efficient influence function $\tilde{\ell}_0 = I_0^{-1}\ell^*$ in the i.i.d. setting is computed from the efficient score

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

and the efficient information

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

for θ in the i.i.d. setting where $M_k(s) = P_{\theta_0,\Lambda_0}[Z^k e^{\theta_0^T Z} I(Y \geq s)]$, $k = 0, 1$.

EXAMPLE 5.4 (Cox model with current status data). Let $T \sim F$ be a failure time, and $Z = (Z_1, Z_2)$ be covariates. Under the case 1 interval censoring [32], we do not observe T but we only know whether an event occurs before an examination time C. We assume sources are formed based on $V = (C, \Delta, Z_2)$ and Z_1 are collected later. The likelihood in the i.i.d. setting is $\ell(\theta, \Lambda) \equiv \delta \log\{1 - e^{-\Lambda(c)\exp(\theta^T z)}\} - (1-\delta)e^{\theta^T z}\Lambda(c)$. The score for θ and Λ is then

$$\dot{\ell}_{\theta,\Lambda}(c, \delta, z) = z\exp\big(\theta^T z\big)\Lambda(c)\big(\delta r(c, z; \theta, \Lambda) - (1-\delta)\big),$$

$$B_{\theta,\Lambda}(h)(c, \delta, z) = \exp\big(\theta^T z\big)h(c)\big\{\delta r(c, z; \theta, \Lambda) - (1-\delta)\big\},$$

where $r(c, z; \theta, \Lambda) = \exp(-e^{\theta^T z}\Lambda(c))/\{1 - \exp(-e^{\theta^T z}\Lambda(c))\}$ (see [31] for details). Our proposed estimator $(\hat{\theta}_N, \hat{\Lambda}_N)$ is the solution to $\mathbb{P}_N^{\mathrm{H}}\dot{\ell}_{\theta,\Lambda} = 0$ and $\mathbb{P}_N^{\mathrm{H}} B_{\theta,\Lambda}(h) = 0$. Conditions 5.4–5.7 can be verified along the same line as in [52] by replacing their weighted empirical process results by our H-empirical process results. In particular, our U-LLN (Theorem 3.1) is used for consistency, and Theorem 5.2 establishes the rate of convergence of $\hat{\Lambda}_N$ as $N^{1/3}$ in view of $(\hat{\theta}_N, \hat{\Lambda}_N)$ as the maximizer of $\mathbb{P}_N^{\mathrm{H}}\ell(\theta, \Lambda)$. This rate agrees with the one in the i.i.d. setting [31]. Then Theorem 5.4 yields

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to_d \mathbb{G}^{\mathrm{H}}\tilde{\ell}_0 \sim N\left(0, I_0^{-1} + \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} \mathrm{Var}_0^{(j)}\left(\rho^{(j)}\tilde{\ell}_0\right)\right).$$

Here, $\tilde{\ell}_0 = I_0^{-1}\ell_0^*$ with

$$\ell_0^* \equiv e^{\theta_0^T z} Q(c, \delta, z; \theta_0, \Lambda_0)\Lambda_0(c)\left\{z - \frac{E[Z e^{2\theta_0^T Z} O(C|Z)|C = c]}{E[e^{2\theta_0^T Z} O(C|Z)|C = c]}\right\}$$

and $I_0 = P_0(\ell_0^*)^{\otimes 2}$ where $Q(c, \delta, z; \theta, \Lambda) = \delta r(c, z; \theta, \Lambda) - (1-\delta)$ and $O(c|z) = \{1 - F(c)\}^{\exp(\theta_0^T z)}/[1 - \{1 - F(c)\}^{\exp(\theta_0^T z)}]$.

6. Numerical results.

6.1. *Simulation studies*. We conducted simulation studies to evaluate finite sample properties of our proposed estimator in the Cox model with right censoring discussed in Example 5.3. Linear and logistic regression models are treated in Section G of the Supplementary Material [51]. Data were generated from the Cox model with two independent covariates $Z_1 \sim \text{Bernoulli}(0.5)$ and $Z_2 \sim N(0,1)$. The failure time T follows $\text{Weibull}(\alpha, \beta), \alpha = 0.2, \beta \in \{0.5, 5\}$ at the baseline and is subject to independent censoring by $C \sim \text{Uniform}(0, c)$ where c was chosen to yield approximately 85% censoring. The regression coefficients are $\theta = (\theta_1, \theta_2)$ with $\theta_1 = \theta_2 \in \{0, \log(1.2), \log(2)\}$. The auxiliary binary variable U is correlated with Z_1 through sensitivity $P(U = 1|Z_1 = 1) = 0.9$ and specificity $P(U = 0|Z_1 = 0) = 0.9$.

We considered four scenarios based on the formation of data sources. Data sources are $\mathcal{V}^{(1)} = \{V : Z_2 \geq -1\}$ and $\mathcal{V}^{(2)} = \{V : Z_2 \leq 1\}$ in Scenario 1 and $\mathcal{V}^{(1)} = \mathcal{V}$ and $\mathcal{V}^{(2)} = \{V : Z_2 \leq 1\}$ in Scenario 2. Sampling fractions in both scenarios were 20% and 30%. In Scenario 3, data sources are $\mathcal{V}^{(1)} = \mathcal{V}$ and $\mathcal{V}^{(2)} = \{V : \Delta = 1\}$ with sampling fractions 20% and 100%. Scenario 4 has three sources where membership in the first two were determined via multinomial logistic regression with Z_2 as a covariate and the third source is $\mathcal{V}^{(3)} = \{V : \Delta = 1\}$. Sampling fractions were 10%, 10% and 100%, respectively. Average sample sizes and numbers of duplications over 2000 datasets in each scenario are shown in Table 1.

The Monte Carlo sample bias and standard deviation of the proposed estimator with ρ from Proposition 3.1 are reported in Table 2. The results show that bias is

TABLE 1
Sample sizes and the numbers of duplications based on 2000 *simulated datasets*

	$\mathcal{V}^{(1)}$	$\mathcal{V}^{(2)}$	N	$N^{(1)}$	$N^{(2)}$	$n^{(1)}$	$n^{(2)}$	Duplication
Scenario 1	$Z_2 \geq -1$	$Z_2 \leq 1$	500	421	421	85	127	21
			10,000	8413	8414	1683	2525	410
Scenario 2	$\mathcal{V}$	$Z_2 \leq 1$	500	500	421	100	127	25
			10,000	10,000	8413	2000	2524	505
Scenario 3	$\mathcal{V}$	$\Delta = 1$	500	500	76	100	76	15
			10,000	10,000	1529	2000	1529	305

								Duplication	
	N	$N^{(1)}$	$N^{(2)}$	$N^{(3)}$	$n^{(1)}$	$n^{(2)}$	$n^{(3)}$	twice	thrice
Scenario 4	500	76	423	278	76	43	28	13	1
	10,000	8475	5564	1529	848	556	1529	258	9

TABLE 2

Results of Monte Carlo simulations with different θ, (α, β), and scenarios

(α, β)		(0.2, 0.5)						(0.2, 5.0)					
θ		log(2)		log(1.2)		0		log(2)		log(1.2)		0	
N		500	10,000	500	10,000	500	10,000	500	10,000	500	10,000	500	10,000
						Complete data (MLE)							
θ_1	Bias	0.004	0.0031	0.004	0.0002	0.001	0.0003	0.017	0.0000	0.001	0.0016	0.001	0.0024
	SD	0.246	0.0534	0.241	0.0531	0.236	0.0518	0.244	0.0530	0.236	0.0522	0.234	0.0518
θ_2	Bias	0.004	0.0004	0.006	0.0005	0.001	0.0008	0.011	0.0020	0.004	0.0002	0.004	0.0004
	SD	0.121	0.0270	0.119	0.0259	0.120	0.0259	0.129	0.0274	0.117	0.0260	0.122	0.0255
						Scenario 1							
θ_1	Bias	0.024	0.0061	0.014	0.0020	0.011	0.0017	0.022	0.0031	0.002	0.0026	0.005	0.0004
	SD	0.482	0.0985	0.432	0.0914	0.429	0.0887	0.477	0.0977	0.435	0.0905	0.423	0.0889
	SEE	0.467	0.0989	0.425	0.0908	0.419	0.0899	0.471	0.0973	0.427	0.0892	0.425	0.0891
θ_2	Bias	0.005	0.0031	0.005	0.0031	0.011	0.0011	0.050	0.0000	0.004	0005	0.008	0.0010
	SD	0.251	0.0526	0.242	0.0486	0.234	0.0495	0.277	0.0544	0.248	0.0496	0.249	0.0505
	SEE	0.260	0.0524	0.248	0.0509	0.244	0.0507	0.285	0.0550	0.254	0.0504	0.254	0.0503
						Scenario 2							
θ_1	Bias	0.062	0.0005	0.017	0.0014	0.009	0.0010	0.034	0.0012	0.019	0.0030	0.008	0.0054
	SD	0.479	0.0967	0.421	0.0894	0.416	0.0876	0.469	0.0941	0.425	0.0889	0.404	0.0899
	SEE	0.467	0.0981	0.421	0.0901	0.412	0.0871	0.459	0.0952	0.423	0.0888	0.415	0.0883
θ_2	Bias	0.016	0.0000	0.003	0.0001	0.015	0.0001	0.026	0.0027	0.002	0.0005	0.010	0.0015
	SD	0.250	0.0526	0.226	0.0499	0.222	0.0493	0.259	0.0510	0.238	0.0486	0.238	0.0501
	SEE	0.252	0.0510	0.238	0.0493	0.232	0.0480	0.267	0.0527	0.241	0.0489	0.236	0.0487

TABLE 2
(*Continued*)

(α, β)		(0.2, 0.5)						(0.2, 5.0)					
θ		log(2)		log(1.2)		0		log(2)		log(1.2)		0	
N		500	10,000	500	10,000	500	10,000	500	10,000	500	10,000	500	10,000
						Scenario 3							
θ_1	Bias	0.005	0.0009	0.008	0.0028	0.006	0.0011	0.025	0.0008	0.008	0.0004	0.006	0.0004
	SD	0.330	0.0733	0.309	0.0660	0.301	0.0676	0.399	0.0860	0.391	0.0840	0.395	0.0841
	SEE	0.330	0.0728	0.308	0.0674	0.305	0.0668	0.375	0.0856	0.365	0.0826	0.366	0.0828
θ_2	Bias	0.023	0.0003	0.018	0.0010	0.001	0.0007	0.029	0.0027	0.016	0.0001	0.001	0.0014
	SD	0.181	0.0378	0.157	0.0341	0.163	0.0342	0.193	0.0437	0.201	0.0422	0.202	0.0414
	SEE	0.171	0.0381	0.156	0.0339	0.156	0.0334	0.183	0.0427	0.181	0.0413	0.183	0.0414
						Scenario 4							
θ_1	Bias	0.010	0.0019	0.003	0.0001	0.005	0.0003	0.060	0.0023	0.016	0.0023	0.002	0.0001
	SD	0.368	0.0789	0.354	0.0760	0.372	0.0775	0.432	0.0970	0.466	0.1031	0.481	0.1022
	SEE	0.355	0.0789	0.343	0.0758	0.347	0.0765	0.401	0.0965	0.417	0.0996	0.418	0.1010
θ_2	Bias	0.023	0.0018	0.006	0.0007	0.012	0.0016	0.011	0.0038	0.013	0.0011	0.037	0.0019
	SD	0.192	0.0407	0.179	0.0363	0.185	0.0367	0.222	0.0477	0.235	0.0489	0.239	0.0499
	SEE	0.181	0.0407	0.169	0.0366	0.169	0.0367	0.198	0.0470	0.203	0.0484	0.202	0.0492

Note: MLE, maximum likelihood estimator based on N items, Bias, an absolute Monte Carlo sample bias; SD, a Monte Carlo sample standard deviation; SEE, average of a plug-in estimator of a standard error.

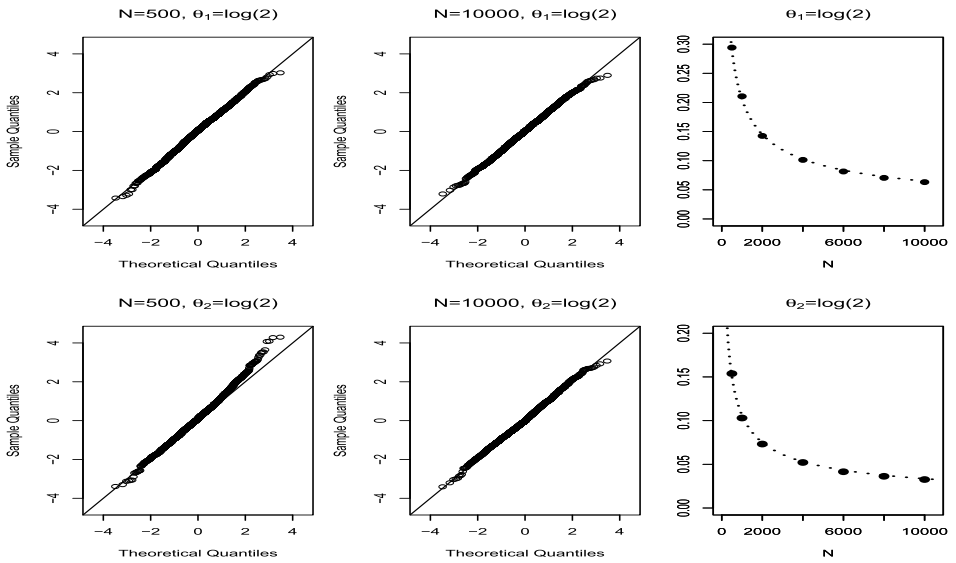

FIG. 2. *Q–Q plots of* $\sqrt{N}(\hat{\theta}-\theta_0)/\widehat{SE}(\hat{\theta})$ *superimposed by* $y=x$ *and plots of averages of absolute differences* $\|\hat{\theta}_N-\theta_0\|$ *against* N *superimposed by* $y=c/x^{1/2}, c=6.5, 3.4$ *in Scenario* 4 *where* $\widehat{SE}(\hat{\theta})$ *is a plug-in estimator of a standard error of* $\sqrt{N}(\hat{\theta}-\theta_0)$.

small and standard deviations are close to averages of plug-in estimators of standard errors in each setting. In Figure 2, right panels show that averages of absolute deviations $\|\hat{\theta}_N-\theta_0\|$ are proportional to $1/N^{1/2}$, and Q–Q plots of the scaled estimators indicate their distributions are approximately the standard normal distribution. Table 3 displays comparison of three different calibration methods in Section 4 and two other choices of ρ [the extension of the single-frame estimator of [35] studied by [41] (SF), and balanced weights of the inverse of the number of sources to which an item belongs (B)] in Scenario 4. Results show that the estimator with the proposed weights ρ and calibration achieved the smallest standard deviations in all cases. All of the above results provide numerical support for our theory. Discussion of additional results and a plug-in variance estimator is provided in Section G of the Supplementary Material [51]. Note that our estimator did not lose much efficiency compared to the MLE for complete data if we base comparison on the number of items used for estimation. For example, 2933 items with duplication were used for our estimator on average when $N=10{,}000$ in Scenario 4 and its standard deviations are 0.0789 and 0.0407 with $(\alpha,\beta)=(0.2,0.5)$ and $\theta_1=\theta_2=\log(2)$. In this case, standard deviations of the MLE based on 2933 items are expected to be about 0.0986 and 0.0499.

6.2. *Real data example.* We illustrate our methods with the national Wilms tumor study (NWTS) [14] where 3915 patients with Wilms tumor were followed until the disease progression. Data contain complete information of all subjects,

TABLE 3
Comparison of calibrations and ρ by standard deviations in Scenario 4

$(\alpha, \beta) = (0.2, 0.5)$	$N = 500$				$N = 10{,}000$			
$\theta_1 = \log(2)$	w/o	SC	C	DC	w/o	SC	C	DC
MLE	0.246				0.0534			
S	0.368	0.333	0.370	0.371	0.0789	0.0720	0.0789	0.0789
SF	0.375	0.341	0.376	0.376	0.0809	0.0740	0.0809	0.0804
B	0.497	0.474	0.497	0.497	0.1060	0.1005	0.1060	0.1060
$\theta_2 = \log(2)$	w/o	SC	C	DC	w/o	SC	C	DC
MLE	0.121				0.0270			
S	0.192	0.188	0.193	0.193	0.0407	0.0395	0.0405	0.0403
SF	0.197	0.192	0.197	0.196	0.0414	0.0401	0.0412	0.0409
B	0.258	0.253	0.258	0.258	0.0530	0.0517	0.0530	0.0530

Note: S, the proposed weights; SF, ρ for a single-frame estimator; B, a balanced weights; w/o, non-calibration; SC, the proposed calibration; C, the standard calibration; DC, the data-source-specific calibration. All calibrations use U and Y.

and was used to compare different stratified designs in [6]. To compare our methods with the MLE with the full cohort and the weighted likelihood estimator with stratified sampling [7], we randomly divided the dataset into two, applied three methods with different designs to training data, and computed the partial likelihood based on testing data. Data sources are deceased subjects, subjects with unfavorable histology measured at hospitals subject to misclassification and the entire cohort with sampling fractions 100%, 50% and 10% resulting in selecting 506 subjects with duplications (438 without duplication). Strata for stratified sampling are deceased subjects, living subjects with unfavorable histology and the rest with sampling fractions 100%, 50% and 14% yielding 502 selected subjects. We fitted data to the Cox model to identify predictors of the relapse of Wilms tumor. Results are summarized in Table 4. The MLE is considered to be most reliable. Estimates from merged and stratified data were all similar to the MLE except the one for cancer stage. Estimated standard errors of the proposed estimator were smaller than those of the estimator with balanced ρ but larger than those from stratified data because stratified sampling effectively used information by avoiding duplication at the design stage. These differences, however, were rather small relative to the magnitudes of estimates even when making comparison with the MLE (except cancer stage). The partial likelihood at the proposed estimator shows better fit than in stratified sampling though the estimator with balanced ρ yielded a larger value. Overall, good performance of the proposed estimator illustrates the potential of data integration as an alternative to stratified sampling.

TABLE 4
Point estimates and estimated standard errors in the analysis of the NWTS study with different sampling schemes

			Data integration					
ρ	**Full cohort**		**Proposed**		**Balanced**		**Stratified sampling**	
# subjects	1957		438 (506 with duplication)				502	
Duplication	0		64 (twice)		2 (thrice)		0	
Partial likelihood	−2458.8		−2464.7		−2463.2		−2467.2	
Variable	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE
Histology	1.430	0.125	1.243	0.236	1.383	0.268	1.419	0.190
Age	0.084	0.021	0.045	0.043	0.043	0.047	0.110	0.035
Stage (III/IV)	1.506	0.356	2.680	0.761	2.589	0.848	2.157	0.705
Tumor	0.064	0.020	0.082	0.046	0.076	0.052	0.106	0.041
Stage × Tumor	−0.079	0.029	−0.156	0.061	−0.079	0.068	−0.134	0.055

Note: Histology is measured at a central laboratory.

7. Discussion. We developed large sample theory for merged data from multiple sources. We proved two limit theorems for the H-empirical process, and applied them to study asymptotic properties of infinite-dimensional M-estimation. Our theory is a nontrivial extension of empirical process theory to a dependent and biased sample with duplication.

We adopted Hartley's estimator as a building block for our theory. This estimator and its variants have been extensively studied under multiple-frame surveys in sampling theory. To conclude this paper, we briefly describe two approaches in sampling theory to illustrate differences from ours.

A primary difference lies in probabilistic frameworks. Sampling theory adopts a finite-population framework where randomness arises only from selection of units. Parameters are finite-population parameters such as sample averages, and statistical models are outside the scope. Asymptotic results usually assume the existence of CLT a priori and asymptotic variance is defined as limits of deterministic sequences (see, e.g., [41, 45, 49, 53, 54]). This difference leads to different optimal ρ and calibration as seen above.

Another less common approach called the super-population framework [27, 34, 50] adopts a similar two-stage formulation [50] but two qualitatively distinct sets of conditions are assumed for different stages of sampling. These conditions concern specific random and nonrandom vectors instead of treating a class of functions in a systematic way. Applications are thus limited to (generalized) linear models [42, 45] where variance estimators (page 4690 of [42], page 1514 of [45]) are our variance estimator for the first stage only. This seeming discrepancy reflects a distinction in probabilistic frameworks.

Acknowledgments. We owe thanks to the Associate Editor and two anonymous referees for their constructive suggestions, which significantly improved the paper. We also thank Jon Wellner for helpful discussions of empirical process theory.

SUPPLEMENTARY MATERIAL

Supplement to "Large sample theory for merged data from multiple sources." (DOI: 10.1214/18-AOS1727SUPP; .pdf). The proofs and additional simulations are given in the Supplement [51].

REFERENCES

[1] ALEXANDER, K. S. (1985). Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 475–493. Wadsworth, Belmont, CA. MR0822047

[2] BAE, J. and LEVENTAL, S. (1995). Uniform CLT for Markov chains and its invariance principle: A martingale approach. *J. Theoret. Probab.* **8** 549–570. MR1340827

[3] BERKES, I. and PHILIPP, W. (1977/78). An almost sure invariance principle for the empirical distribution function of mixing random variables. *Z. Wahrsch. Verw. Gebiete* **41** 115–137. MR0464344

[4] BERTAIL, P., CHAUTRU, E. and CLÉMENÇON, S. (2017). Empirical processes in survey sampling with (conditional) Poisson designs. *Scand. J. Stat.* **44** 97–111. MR3619696

[5] BOISTARD, H., LOPUHAÄ, H. P. and RUIZ-GAZEN, A. (2017). Functional central limit theorems for single-stage sampling designs. *Ann. Statist.* **45** 1728–1758. MR3670194

[6] BRESLOW, N. E. and 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** 457–468.

[7] BRESLOW, N. E. and WELLNER, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Stat.* **34** 86–102. MR2325244

[8] BRESLOW, N. E. and WELLNER, J. A. (2008). A Z-theorem with estimated nuisance parameters and correction note for: "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression" [Scand. J. Statist. **34** (2007) 86–102; MR2325244]. *Scand. J. Stat.* **35** 186–192. MR2391566

[9] BRICK, J. M., DIPKO, S., PRESSER, S., TUCKER, C. and YUAN, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opin. Q.* **70** 780–793.

[10] CANTELLI, F. P. (1933). Sulla determinazione empirica delle leggi di probabilita. *G. Ist. Ital. Attuari* **4** 421–424.

[11] CERVANTES, I. F., JONES, M. E., ROJAS, L. A., BRICK, J. M., KURATA, J. and GRANT, D. (2006). A review of the sample design for the California health interview survey. In *Proceedings of the Social Statistics Section* 3023–3030. Amer. Statist. Assoc., Alexandria, VA.

[12] CHATTERJEE, N., CHEN, Y.-H., MAAS, P. and 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** 107–117. MR3494641

[13] COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

[14] D'ANGIO, G. J., BRESLOW, N., BECKWITH, J. B., EVANS, A., BAUM, H., DELORIMIER, A., FERNBACH, D., HRABOVSKY, E., JONES, B. and KELALIS, P. (1989). Treatment of Wilms' tumor. Results of the Third National Wilms' Tumor Study. *Cancer* **64** 349–360.

[15] DE LEEUW, E. D. (2005). To mix or not to mix data collection modes in surveys. *J. Off. Stat.* **21** 233–255.

[16] DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

[17] DILLMAN, D. A., SMYTH, J. D., CHRISTIAN and MELANI, L. (2014). *Internet*, *Phone*, *Mail*, *and Mixed-Mode Surveys*: *The Tailored Design Method*, 4th ed. Wiley, New York.

[18] DING, Y. and NAN, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Ann. Statist.* **39** 3032–3061. MR3012400

[19] DONSKER, M. D. (1952). Justification and extension of Doob's heuristic approach to the Komogorov–Smirnov theorems. *Ann. Math. Stat.* **23** 277–281. MR0047288

[20] DUDLEY, R. M. (1981). Donsker classes of functions. In *Statistics and Related Topics* (*Ottawa*, *Ont*., 1980) 341–352. North-Holland, Amsterdam. MR0665285

[21] FELLEGI, I. P. and SUNTER, A. B. (1969). A theory for record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.

[22] GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12** 929–998. MR0757767

[23] GLIVENKO, V. (1933). Sulla determinazione empirica della legge di probabilita. *G. Ist. Ital. Attuari* **4** 92–99.

[24] HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 361–374. MR0125612

[25] HARTLEY, H. O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section* 203–206. Amer. Statist. Assoc., Alexandria,VA.

[26] HARTLEY, H. O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, *Ser. C* **36** 99–118.

[27] HARTLEY, H. O. and SIELKEN, R. L. JR. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics* **31** 411–422. MR0386084

[28] HERZOG, T. N., SCHEUREN, F. J. and WINKLER, W. E. (2007). *Data Quality and Record Linkage Techniques*. 1st ed. Springer, Berlin.

[29] HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

[30] HU, S. S., BALLUZ, L., BATTAGLIA, M. P. and FRANKEL, M. R. (2011). Improving public health surveillance using a dual-frame survey of landline and cell phone numbers. *Am. J. Epidemiol.* **173** 703–711.

[31] HUANG, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24** 540–568. MR1394975

[32] HUANG, J. and WELLNER, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*: *Survival Analysis* 123–169. Springer, Berlin.

[33] IACHAN, R. and DENNIS, M. L. (1993). A multiple frame approach to sampling the homeless and transient population. *J. Off. Stat.* **9** 747–764.

[34] ISAKI, C. T. and FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77** 89–96. MR0648029

[35] KALTON, G. and ANDERSON, D. W. (1986). Sampling rare populations. *J. R. Stat. Soc., A* **149** 65–82.

[36] KEIDING, N. and LOUIS, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. Roy. Statist. Soc. Ser. A* **179** 319–376. MR3461587

[37] KIM, G. and CHAMBERS, R. (2012). Regression analysis under incomplete linkage. *Comput. Statist. Data Anal.* **56** 2756–2770. MR2915160
[38] KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. Springer, New York. MR2724368
[39] LAHIRI, P. and LARSEN, M. D. (2005). Regression analysis with linked data. *J. Amer. Statist. Assoc.* **100** 222–230. MR2156832
[40] LEVENTAL, S. (1989). A uniform CLT for uniformly bounded families of martingale differences. *J. Theoret. Probab.* **2** 271–287. MR0996990
[41] LOHR, S. and RAO, J. N. K. (2006). Estimation in multiple-frame surveys. *J. Amer. Statist. Assoc.* **101** 1019–1030. MR2324141
[42] LU, Y. (2012). Regression coefficient estimation in dual frame surveys. In *Proceedings of the Section on Survey Research Methods* 4687–4695. Amer. Statist. Assoc., Alexandria, VA.
[43] LU, Y. and LOHR, S. L. (2010). Gross flow estimation in dual frame surveys. *Surv. Methodol.* **36** 13–22.
[44] MA, S. and KOSOROK, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.* **96** 190–217. MR2202406
[45] METCALF, P. and SCOTT, A. (2009). Using multiple frames in health surveys. *Stat. Med.* **28** 1512–1523. MR2649709
[46] PRÆSTGAARD, J. and WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301
[47] RANALLI, M. G., ARCOS, A., DEL MAR RUEDA, M. and TEODORO, A. (2016). Calibration estimation in dual-frame surveys. *Stat. Methods Appl.* **25** 321–349. MR3539496
[48] RAO, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Off. Stat.* **10** 153–165.
[49] RAO, J. N. K. and WU, C. (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *J. Amer. Statist. Assoc.* **105** 1494–1503. MR2796566
[50] RUBIN-BLEUER, S. and SCHIOPU KRATINA, I. (2005). On the two-phase framework for joint model and design-based inference. *Ann. Statist.* **33** 2789–2810. MR2253102
[51] SAEGUSA, T. (2019). Supplement to "Large sample theory for merged data from multiple sources." DOI:10.1214/18-AOS1727SUPP.
[52] SAEGUSA, T. and WELLNER, J. A. (2013). Weighted likelihood estimation under two-phase sampling. *Ann. Statist.* **41** 269–295. MR3059418
[53] SKINNER, C. J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *J. Amer. Statist. Assoc.* **86** 779–784. MR1147105
[54] SKINNER, C. J. and RAO, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *J. Amer. Statist. Assoc.* **91** 349–356. MR1394091
[55] VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1999). *Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. MR1915446
[56] VAN DER VAART, A. W. (1995). Efficiency of infinite-dimensional M-estimators. *Stat. Neerl.* **49** 9–30. MR1333176
[57] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247
[58] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. Springer, New York. MR1385671
[59] WINKLER, W. E. (1995). *Matching and Record Linkage* 353–384. Wiley, New York.
[60] ZIEGLER, K. (1997). Functional central limit theorems for triangular arrays of function-indexed processes under uniformly integrable entropy conditions. *J. Multivariate Anal.* **62** 233–272. MR1473875
[61] ZIEGLER, K. (2001). Uniform laws of large numbers for triangular arrays of function-indexed processes under random entropy conditions. *Results Math.* **39** 374–389. MR1834583

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
4303 KIRWAN HALL
COLLEGE PARK, MARYLAND 20742
USA
E-MAIL: tsaegusa@math.umd.edu