PATIENT RECRUITMENT USING ELECTRONIC HEALTH RECORDS UNDER SELECTION BIAS: A TWO-PHASE SAMPLING FRAMEWORK

BY GUANGHAO ZHANG[1,*], LAUREN J. BEESLEY[2], BHRAMAR MUKHERJEE[1,†], AND XU SHI[1,‡]

[1]*Department of Biostatistics, University of Michigan,* [*]*ghzhang@umich.edu;* [†]*bhramar@umich.edu;* [‡]*shixu@umich.edu*

[2]*Statistical Sciences Group, Los Alamos National Laboratory, lvandervort@lanl.gov*

Electronic health records (EHRs) are increasingly recognized as a cost-effective resource for patient recruitment in clinical research. However, how to optimally select a cohort from millions of individuals to answer a scientific question of interest remains unclear. Consider a study to estimate the mean or mean difference of an expensive outcome. Inexpensive auxiliary covariates predictive of the outcome may often be available in patients' health records, presenting an opportunity to recruit patients selectively which may improve efficiency in downstream analyses. In this paper, we propose a two-phase sampling design that leverages available information on auxiliary covariates in EHR data. A key challenge in using EHR data for multi-phase sampling is the potential selection bias, because EHR data are not necessarily representative of the target population. Extending existing literature on two-phase sampling design, we derive an optimal two-phase sampling method that improves efficiency over random sampling while accounting for the potential selection bias in EHR data. We demonstrate the efficiency gain from our sampling design via simulation studies and an application to evaluating the prevalence of hypertension among US adults leveraging data from the Michigan Genomics Initiative, a longitudinal biorepository in Michigan Medicine.

1. Introduction. Electronic health record (EHR) data are increasingly used to facilitate patient recruitment (Effoe et al., 2016; Cowie et al., 2017; McCord and Hemkens, 2019). An EHR is a digital repository of routinely collected patient health information, including medical diagnosis, procedure, medication, radiology images, and laboratory test (Häyrinen, Saranto and Nykänen, 2008; Shortreed et al., 2019). In conventional observational and randomized studies, patient recruitment and patient retention are often limited by funding and time. In addition, the recruited cohort tends to be homogeneous and not representative of a real-world population, thus study results are often not generalizable to a target population for health policy decision making (Hemkens, Contopoulos-Ioannidis and Ioannidis, 2016). In contrast, the rich clinical information and real-world population provided by EHR present a cost-effective data source to identify and recruit patients who satisfy the eligibility criteria of a research study (Bower et al., 2017; Schreiweis et al., 2014). For example, Wu et al. (2017) developed a semantic search system that is used to recruit patients into the $100,000$ Genomes Project leveraging clinical notes in EHR; Thadani et al. (2009) deployed an electronic screening method to identify eligible patients for trial recruitment. However, patient recruitment using EHR data has been limited to random sampling after applying the inclusion/exclusion criteria. Because EHR data from a healthcare system may be biased towards certain demographic groups or specific health conditions, random sampling may lead to a biased cohort. In addition, random sampling may not sufficiently capture a rare event and does not utilize information on risk factors recorded in EHR. It remains unclear how to optimally select a cohort from millions of individuals to answer a scientific question of interest.

Keywords and phrases: auxiliary information, electronic health records, selection bias, study design, two-phase sampling.

The goal of this paper is to improve the usability of EHR data for patient recruitment and population-level parameter estimation with an efficient sampling design framework. A motivating example is the Michigan Genomics Initiative (MGI), a longitudinal biorepository at the University of Michigan Health System that was launched in 2012, linking patient EHR data with genetic data to facilitate biomedical research. Patients 18 years of age or older who underwent surgery at the University of Michigan Health System are approached for enrollment. Participants provide broad opt-in consent for use of their EHR data and biospecimen, as well as recontact in the future for any applicable research study. As such, an increasing number of survey, experimental, and observational studies have been conducted by recruiting patients from the MGI cohort based on their medical records (Joyce et al., 2021; Wu et al., 2021). For example, a sample of eligible MGI patients were surveyed in March and April 2020 to evaluate risk factors for COVID-19 and the impact of the 'Stay Home Stay Safe' executive order on Michigan residents (Wu et al., 2021).

With the growing availability of MGI participants, it is possible to selectively recruit eligible patients to obtain maximal information under a budget constraint. We illustrate this by estimating the prevalence of a chronic disease, such as hypertension, in the US adult population using MGI data. Typically outcome assessment is expensive or time-consuming, while inexpensive auxiliary covariates predictive of the outcome are readily available in EHR data. Therefore, instead of randomly recruiting eligible patients, investigators can leverage such auxiliary information to sample patients from MGI according to an optimized sampling probability to improve efficiency in downstream analyses.

Our proposal is motivated by the observation that patient recruitment using EHR data constitutes a two-phase sampling framework. As illustrated in Figure 1, from a pre-specified target population, a subset of individuals seek healthcare and their information is recorded in an EHR system. We refer to such a cohort as the EHR sample or the phase-I sample. From the phase-I EHR sample, we aim to recruit a subset of patients into the study sample, i.e., the phase-II sample in a way such that the ultimate results are generalizable to a target population of inference (e.g., the US adult population). A review of the two-phase sampling literature is provided in Section 1 of the Supplementary Material (Zhang et al., 2022a).

However, existing two-phase sampling methods cannot be directly applied because the phase-I EHR sample such as the MGI cohort is not necessarily a random sample of the target population (Phelan, Bhavsar and Goldstein, 2017; Goldstein et al., 2016; Beesley et al., 2020). Patients are observed in EHR data only if they seek care. When the EHR sample differs from the target population in terms of characteristics relevant to the scientific question in view, parameter estimates could be biased and statistical conclusions may lack generalizability (Tripepi et al., 2010). Several methods to model and mitigate selection bias in EHR data have been developed recently. For example, Haneuse and Daniels (2016) proposed to model the selection mechanism by breaking it down into submechanisms due to a sequence of decisions made by patients, providers, and healthcare systems. Goldstein et al. (2016) considered addressing selection bias in EHR by controlling for healthcare utilization measured by number of medical encounters. Beesley and Mukherjee (2020) proposed calibration weighting and inverse probability of selection weighting methods to account for the differences between patients included and not included in the EHR cohort.

In this paper, we propose an optimal design for patient recruitment using EHR data under a two-phase sampling framework while accounting for potential selection bias in the EHR cohort. We first present an estimator of the mean outcome that acknowledges selection bias and leverages auxiliary information in EHR data. We then derive the optimal sampling probability for recruiting patients from the EHR cohort into the study cohort which minimizes the variance of this estimator. We extend existing two-phase sampling methods to further account for selection bias with two approaches: direct estimation of selection mechanism and indirect

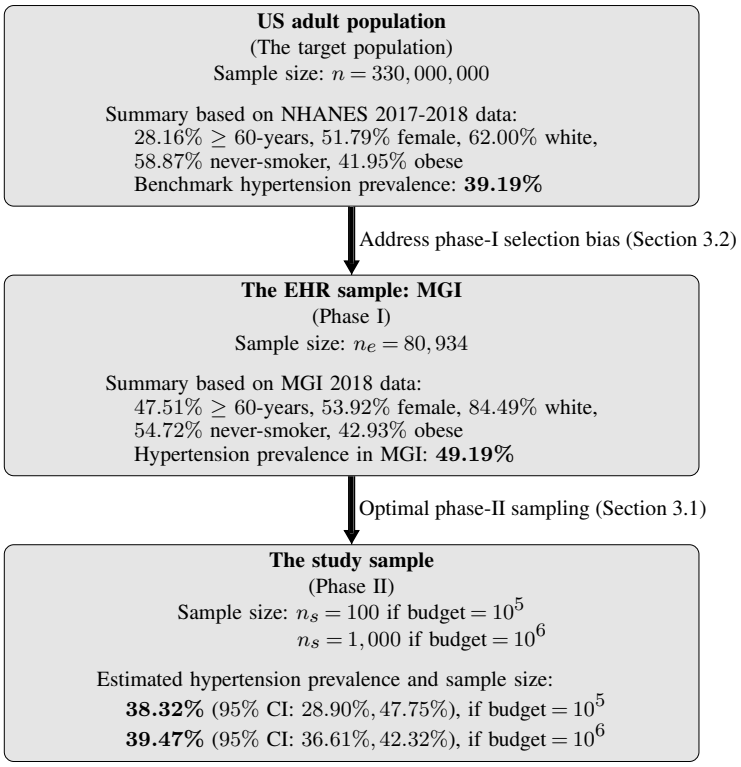

FIG 1. *The two-phase sampling framework for selective patient recruitment using EHR data from the MGI: illustrated by estimating hypertension prevalence in the US adult population. Hypertension is chosen as the outcome of interest because it is readily measured in MGI and NHANES data which allows for validation. The raw prevalence in MGI (49.19%) is much higher than the benchmark value computed based on NHANES (39.19%), indicating selection bias in the EHR sample. We aim to optimally recruit patients from the MGI cohort into a study sample to improve efficiency in estimation of hypertension prevalence while accounting for potential selection bias.*

bias reduction via subsampling strategy. We ultimately provide a two-phase design framework for estimation of a general estimand that is the solution to a given estimating equation, such as the average treatment effect (ATE) when the phase-II study is either an observational study or a randomized controlled trial (RCT), and coefficients in a regression model.

The rest of the paper is organized as follows. Section 2 formulates EHR-based patient recruitment into a two-phase sampling framework and presents design and estimation strategies under this framework. We detail our proposed methods in Section 3, where we derive the optimal phase-II sampling probability under a given budget allowing for a biased phase-I EHR sample in Section 3.1, present two methods for addressing the selection bias in EHR data in Section 3.2, and detail an estimator of the parameter of interest in Section 3.3. We prove the efficiency gain compared to random sampling in Section 4. In Section 5, we extend our proposed framework to facilitate estimation of a general estimand such as the ATE and regression coefficients. We demonstrate the efficiency gain via extensive simulation studies in Section 6, and we illustrate our proposed methods with an application to estimating the prevalence of hypertension in the US adult population using MGI data in Section 7. We conclude with a brief discussion in Section 8.

2. Preliminaries.

2.1. *The EHR-based two-phase sampling framework.* Let Y denote the outcome of interest that is expensive or time-consuming to measure. Suppose one is interested in estimating

the mean outcome, $\beta = E[Y]$, in the target population. This can be generalized to other parameters of interest, such as the ATE and regression coefficients, which we detail in Section 5. Our goal is to develop an optimal sampling mechanism to improve efficiency in the estimation of β by leveraging the auxiliary information that is inexpensive and readily available in the EHR sample. Below we summarize the samples at each phase, the relationship between the samples, and the data missingness pattern, which are visualized in Figure 2.

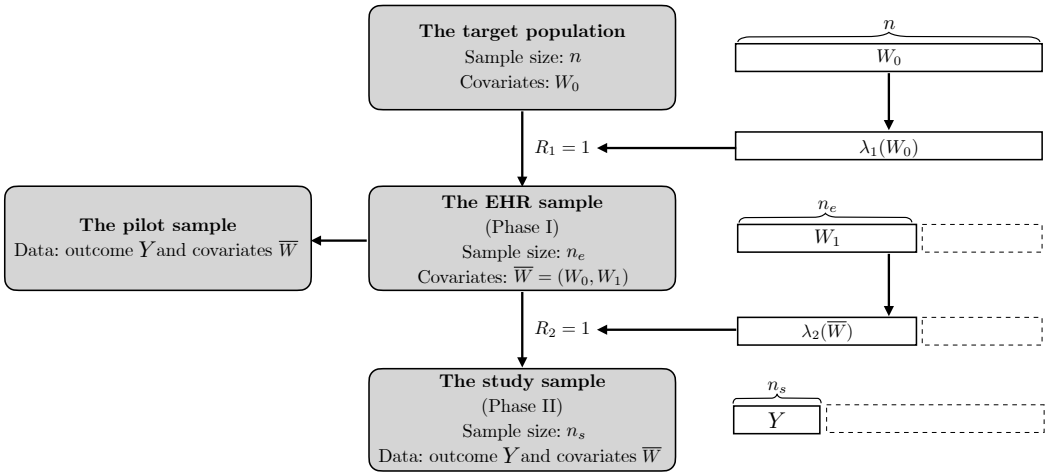

FIG 2. *Flowchart of the EHR-based two-phase sampling framework. We aim to recruit a study sample based on a designed and optimized sampling probability $\lambda_2(\overline{W})$ to minimize the variance of the estimated parameter. The dashed boxes indicate missing data, and the solid boxes indicate observed data at each phase.*

The target population. Consider a target population of n individuals. Let W_0 denote patient characteristics predictive of both the outcome and whether a patient seeks healthcare and thus shows up in the EHR system. We assume that information about W_0 is available in either of the following two scenarios: (1) W_0 is available in an external probability sample, such as a survey study, in which individuals are sampled from the target population with known sampling mechanism; (2) the distribution (or at least summary statistics) of W_0 in the target population is available from a public data source, such as the Census.

The phase-I sample: the EHR sample. Suppose one has access to EHR data on a subset of the target population with sample size $n_e \leq n$, referred to as the EHR sample or the phase-I sample. Let $R_1 \in \{0, 1\}$ be the indicator of whether an individual in the target population is selected into the phase-I sample, then $n_e = \sum_{i=1}^{n} R_{1i}$. The mechanism of being selected into the phase-I sample, i.e., the probability of having EHR available for an individual in the target population, depends on patient characteristics W_0 via an unknown selection mechanism

$$\lambda_1(W_0) = P(R_1 = 1 \mid W_0).$$

Let W_1 denote auxiliary covariates that are available in phase-I in addition to W_0 and are predictive of the outcome. To ease exposition, let $\overline{W} = (W_0, W_1)$ denote all covariates available in the phase-I sample, that is, $\overline{W}_i$ is observed if $R_{1i} = 1, i = 1, \ldots, n$.

The phase-II sample: the study sample. The study sample of size n_s $(1 \leq n_s \leq n_e)$, also referred to as the phase-II sample, will be recruited from the phase-I EHR sample based on patient characteristics $\overline{W}$. Then a clinical outcome of interest, Y, will be measured, which often involves labor-intensive manual chart review in EHR-based research. Let $R_2 \in \{0,1\}$ be the indicator of whether an individual is recruited into the phase-II sample, then $n_s = \sum_{i=1}^{n} R_{2i}$. The probability of being selected into the phase-II sample from the phase-I sample depends on patient characteristics $\overline{W}$ via a designed probability

$$\lambda_2(\overline{W}) = P(R_2 = 1 \mid \overline{W}, R_1 = 1).$$

To summarize, ideally we wish to have i.i.d. sample of the complete data (Y, W_1, W_0), which arises from some joint distribution $P \in \mathcal{M}$ in the target population. Sampling into EHR generates incomplete data $(W_1 R_1, W_0) \sim P_{\text{I}} \in \mathcal{M}_{\text{I}} = \{P_{P,\lambda_1(W_0)} : P \in \mathcal{M}, \lambda_1(W_0)\}$, where model $\mathcal{M}_{\text{I}}$ is implied by biased sampling through $\lambda_1(W_0)$ (i.e., phase-I sampling) from $\mathcal{M}$. To study $\beta = E[Y]$, one then draws a phase-II sample to measure the outcome through $\lambda_2(\overline{W})$, which generates the final observed data $O = (YR_2, W_1R_1, W_0) \sim P_{\text{II}} \in \mathcal{M}_{\text{II}} = \{P_{P_{\text{I}},\lambda_2(\overline{W})} : P_{\text{I}} \in \mathcal{M}_{\text{I}}, \lambda_2(\overline{W})\}$. We aim to find the optimal sampling design, denoted as $\lambda_2^*(\overline{W})$, that is most efficient in the sense that the sampling design minimizes the asymptotic variance for a given estimator of β.

The pilot sample. Design of a study typically relies on either certain prior knowledge or preliminary data obtained from a pilot study. We assume that a small pilot sample of size n_p is available with both Y and $\overline{W}$ measured. The pilot data allow us to model the relationship between the outcome and the auxiliary covariates to inform phase-II sampling design. Ideally, the pilot sample should be a random sample of the phase-I EHR sample. When such a pilot sample is unavailable, knowledge about the conditional variance of the outcome, $Var(Y \mid \overline{W}, R_1 = 1)$, is required.

In addition to the multi-phase samples, we introduce some notation for budget consideration. Let B denote the total budget, C_0 denote the initial study cost, and C_1 denote the per-individual cost in phase-I that scales with the EHR sample size and does not depend on patient characteristics. Let $C_2(\overline{W})$ be the per-individual cost in phase-II that may depend on auxiliary covariates for patient characteristics. For example, in HIV vaccine trials, the cost of measuring immune response is associated with the number of reactive HIV epitopes (Zolla-Pazner, 2004; Sahay, Nguyen and Yamamoto, 2017). The total expected cost, which is constrained to not exceed the total budget B, is given by $C_0 + n_e C_1 + nE\{\lambda_1(W_0)\lambda_2(\overline{W})C_2(\overline{W})\}$, which is derived in Section 2.1 of the Supplementary Material (Zhang et al., 2022a).

2.2. *Estimation under two-phase sampling allowing for a biased phase-I sample.* Gilbert, Yu and Rotnitzky (2014) proposed an optimal phase-II sampling design that minimizes the variance of an estimator of β with a budget constraint, under the assumption that the phase-I sample is a simple random sample of the target population, that is, $W_0 = \emptyset$ and $\lambda_1(W_0)$ is a constant. In our setting, the phase-I sample is a cohort of patients whose EHR data are available to the investigator. The phase-I EHR sample is not necessarily a random sample representative of the target population, but rather is potentially subject to selection bias, which needs to be accounted for. In this section, we extend the optimal sampling design proposed by Gilbert, Yu and Rotnitzky (2014) to account for selection bias in the phase-I EHR sample with an unknown mechanism $\lambda_1(W_0)$.

We impose the following standard assumptions in the missing data literature:

1. Missing at random: $R_2 \perp\!\!\!\perp Y \mid R_1, \overline{W}$ and $R_1 \perp\!\!\!\perp (Y, W_1) \mid W_0$;
2. Positivity: $\epsilon < \lambda_1(W_0), \lambda_2(\overline{W}) \leq 1$ for some $\epsilon > 0$;
3. One of the following two sets of models is correctly specified:

(i) Selection probabilities: $\lambda_1(W_0)$ and $\lambda_2(\overline{W})$;
(ii) Outcome regression model: $E(Y \mid \overline{W}, R_2 = 1)$ and $f(W_1 \mid W_0, R_1 = 1)$ (thus $E(Y \mid W_0, R_1 = 1)$ is also correctly specified).

Assumption 1 implies that in each phase, the sampling mechanism only depends on some auxiliary covariates of the previous phase and does not depend on the outcome and auxiliary covariates of the next phase. Specifically, in phase-I, we assume that the sampling probability is an unknown function of W_0, which is a more flexible setting than the traditional two-phase sampling framework proposed by Gilbert, Yu and Rotnitzky (2014) where $\lambda_1(W_0)$ is assumed to be a constant. In phase-II, Assumption 1 is guaranteed to hold because we will sample patients based on the derived optimal sampling probability $\lambda_2^*(\overline{W})$, which does not depend on Y. In the motivating example of patient recruitment through MGI, auxiliary covariates predictive of the outcome and available in MGI data can serve as $\overline{W}$ in practice. Among $\overline{W}$, covariates with a differential distribution between the MGI sample and the target population should be taken as W_0 to standardize results to the target population.

Assumption 2 requires that individuals with any given value of W_0 (or $\overline{W}$) have nonzero probabilities of being selected into the EHR sample (or study sample). Positivity of $\lambda_2(\overline{W})$ holds by design. The positivity assumption of $\lambda_1(W_0)$ is more stringent. For example, suppose health insurance coverage impacts whether a patient seeks care at the University of Michigan Health System, such that there is zero probability for patients without insurance to show up in MGI data. Then the positivity assumption requires that health insurance coverage as a covariate is not needed for Assumption 1 to hold, that is, access to insurance does not impact the disease outcome. Under a weaker version of Assumption 1 referred to as the mean exchangeability (Shi, Pan and Miao, 2023), i.e., $E[Y \mid W_0, R_1 = 1] = E[Y \mid W_0]$, the positivity assumption requires that health insurance coverage does not modify the outcome model in the MGI sample compared to the target population. In practice, one should carefully evaluate factors that might impact both the outcome of interest and the selection into the EHR sample before applying our proposed method.

Assumption 3 is imposed to ensure consistent estimation of β, as detailed in Section 2.2 of the Supplementary Material (Zhang et al., 2022a).

As illustrated in Figure 2, the EHR-based two-phase sampling framework constitutes a monotone missing data pattern, with independent and identically distributed observations $O_i = (Y_i R_{2i}, W_{1i} R_{1i}, W_{0i})$, $i = 1, \ldots, n$. Under Assumptions 1-3, Rotnitzky and Robins (1995) proposed to estimate β by solving $n^{-1} \sum_{i=1}^{n} U(O_i; \beta) = 0$, where

$$
\begin{aligned}
U(O;\beta) = {} & \frac{R_1 R_2}{\lambda_1(W_0)\lambda_2(\overline{W})} Y - \frac{R_1 R_2 - R_1 \lambda_2(\overline{W})}{\lambda_1(W_0)\lambda_2(\overline{W})} E(Y \mid \overline{W}, R_1 = 1) \\
& - \frac{R_1 - \lambda_1(W_0)}{\lambda_1(W_0)} E(Y \mid W_0) - \beta.
\end{aligned} \tag{1}
$$

We refer to the resulting estimator as the RR estimator hereafter (Rotnitzky and Robins, 1995). The estimating function $U(O;\beta)$ is originally derived for mean outcome estimation in the presence of nonresponse in longitudinal studies. The estimation strategy directly applies to our setting where the sampling mechanisms in multiple phases (see Figure 2) correspond to the nonresponse mechanisms over time. Intuitively, one could estimate β while accounting for biased sampling by inverse probability weighting using the measured outcome from the phase-II sample, which corresponds to the first term, $R_1 R_2 Y / \{\lambda_1(W_0)\lambda_2(\overline{W})\}$. Efficiency can be improved by further incorporating information from the auxiliary covariates through two augmentation terms that essentially imputes missing outcomes using $E(Y \mid W_0)$ and $E(Y \mid \overline{W}, R_1 = 1)$ when the outcome models are correctly specified.

The estimator is doubly robust in the sense that $E(U(O;\beta)) = 0$ when selection probabilities (i.e., $\lambda_1(W_0)$ and $\lambda_2(\overline{W})$) or outcome models (i.e., $E(Y \mid \overline{W}, R_1 = 1)$ and $E(Y \mid W_0)$) are correctly specified, which is proved in Section 2.2 of the Supplementary Material. Based on Assumption (1), we have

$$E(Y \mid \overline{W}, R_1 = 1) = E(Y \mid \overline{W}, R_2 = 1) \text{ and}$$
$$E(Y \mid W_0) = E(Y \mid W_0, R_1 = 1) = E\{E(Y \mid \overline{W}, R_2 = 1) \mid W_0, R_1 = 1\}.$$

Therefore, correct specification of $E(Y \mid \overline{W}, R_1 = 1)$ is equivalent to correct specification of $E(Y \mid \overline{W}, R_2 = 1)$. Similarly, correct specification of $E(Y \mid W_0)$ is equivalent to correct specification of $E(Y \mid \overline{W}, R_2 = 1)$ and $f(W_1 \mid W_0, R_1 = 1)$. Thus consistent estimation is ensured when one of the following two sets of models is correctly specified:

(i) Selection probabilities: $\lambda_1(W_0)$ and $\lambda_2(\overline{W})$;
(ii) Outcome regression model: $E(Y \mid \overline{W}, R_2 = 1)$ and $f(W_1 \mid W_0, R_1 = 1)$.

The asymptotic variance of the RR estimator is given by $V(\lambda_2)/n$, where

$$V(\lambda_2) = Var(Y) + E\left[\left\{\frac{1}{\lambda_1(W_0)} - 1\right\} Var(Y \mid W_0)\right] + E\left[\left\{\frac{1}{\lambda_2(\overline{W})} - 1\right\} \frac{1}{\lambda_1(W_0)} Var(Y \mid \overline{W}, R_1 = 1)\right].$$

3. Methods.

3.1. *Optimal two-phase sampling allowing for a biased phase-I sample.* We aim to find the optimal sampling probability $\lambda_2^*(\overline{W})$ that minimizes the above asymptotic variance of the RR estimator under the constraint that the phase-II sampling probability is non-negative and that the budget covers the total expected cost. This can be formulated as the following optimization problem:

$$\begin{aligned} \arg\min_{\lambda_2(\overline{w})} \ & \frac{V(\lambda_2)}{n} \\ \text{s.t. } & 0 < \lambda_2(\overline{W}) \leq 1 \\ & C_0 + n_e C_1 + nE\{\lambda_1(W_0)\lambda_2(\overline{W})C_2(\overline{W}) - B = 0, \end{aligned}$$

which is a convex optimization problem that can be solved with Karush–Kuhn–Tucker (KKT) conditions. The rationale behind using the KKT conditions is the following. To find the extremum (maximum or minimum value) of a function in an unconstrained optimization problem, one usually searches for an optimal point where the slope or gradient is zero. When the optimization problem is subject to an equality constraint, the use of the Lagrange multiplier helps convert an optimization problem into a system of equations. The solution to the system of equations is the optimal point. The KKT method generalizes the method of Lagrange multipliers to inequality constraints. It has been shown that for a convex optimization problem, the point that satisfies the KKT conditions is the sufficient and necessary solution for optimality (Boyd and Vandenberghe, 2004).

Given a fixed budget B, target population sample size n, and EHR sample size n_e, we now present the optimal sampling probability $\lambda_2^*(\overline{W})$ that satisfies the KKT conditions and hence achieves the minimal variance, which is proved in Section 2.3 of the Supplementary Material (Zhang et al., 2022a).

THEOREM 3.1. *For fixed B and n_e that satisfy $n_e < (B - C_0)/C_1$, the minimal variance among all possible designs that do not exceed the budget B is achieved at*

$$\lambda_2^*(\overline{W} = \overline{w}) = \min\left\{1, \frac{B - C_0 - n_e C_1}{n\lambda_1(w_0)} \frac{\sqrt{Var(Y \mid \overline{W}, R_1 = 1)/C_2(\overline{w})}}{E\left\{\sqrt{C_2(\overline{W})Var(Y \mid \overline{W}, R_1 = 1)}\right\}}\right\}. \tag{2}$$

The upper bound for n_e is to make sure that we have enough budget to cover the per-individual cost. We can see that $\lambda_2^*(\overline{W})$ increases with $Var(Y \mid \overline{W}, R_1 = 1)/C_2(\overline{W})$, which is the cost-standardized conditional variance of Y. Intuitively, individuals with noninformative auxiliary covariates (i.e., $Var(Y \mid \overline{W}, R_1 = 1)$ is large) and relatively affordable outcome measurement costs (i.e., $C_2(\overline{W})$ is low) will be oversampled by the proposed design, because for these individuals, it might be more efficient to measure their outcome Y directly. Conversely, individuals with informative auxiliary covariates and expensive outcome measurement costs will be undersampled, because for these individuals, instead of directly measuring the outcome, it is more efficient to impute the unobserved outcome leveraging the highly predictive auxiliary covariates.

It is important to note that $Var(Y \mid \overline{W}, R_1 = 1)$ and $\lambda_1(W_0)$ in Eq. (2) are generally unknown and need to be estimated to inform phase-II sampling. We propose to estimate $Var(Y \mid \overline{W}, R_1 = 1)$ using data from the pilot sample, which is detailed in Section 2.4 of the Supplementary Material (Zhang et al., 2022a). When individual-level data on W_0 are available in the target population, it is straightforward to estimate $\lambda_1(W_0)$ by running a regression model for $P(R_1 = 1 \mid W_0)$. However, in practice, we generally do not have individual-level data from the target population. We address this issue in the next section.

3.2. *Addressing selection bias in the phase-I EHR sample.* In this section, we present two methods to account for potential selection bias $\lambda_1(W_0)$ in the EHR sample: direct estimation of $\lambda_1(W_0)$ and indirect bias reduction via subsampling.

3.2.1. *Method 1: direct estimation of selection mechanism.* Beesley and Mukherjee (2020) proposed novel strategies to model $\lambda_1(W_0)$ by leveraging an external probability sample of the target population with both sampling probability and the auxiliary covariates available, assuming there is no overlap between the external probability sample and the EHR sample. Such external data are often available in survey studies. For example, one can use the publicly available National Health and Nutrition Examination Survey (NHANES) data as the external probability sample when the target population is defined as the US adult population. The rationale is to approximate the phase-I sampling probability by calibrating the sampling probability of the external probability sample (such as NHANES) (Elliot, 2013).

We employ the method proposed by Beesley and Mukherjee (2020) to estimate the probability of being selected into the phase-I EHR sample. Specifically, we have that

$$\lambda_1(W_0) = P(R_1 = 1 \mid W_0) \approx P(R_{\text{prob}} = 1 \mid W_0) \frac{P(R_1 = 1 \mid W_0, R_{\text{comb}} = 1)}{1 - P(R_1 = 1 \mid W_0, R_{\text{comb}} = 1)}, \tag{3}$$

where R_{prob} is the indicator of being included in the external probability sample from the target population, R_{comb} is the indicator of being included in the sample combining both the EHR sample and the external probability sample, that is, $R_{\text{comb}} = \mathbb{I}(R_{\text{prob}} = 1 \text{ or } R_1 = 1)$.

To estimate $P(R_{\text{prob}} = 1 \mid W_0)$, we take the sampling probability in the external probability sample as a continuous outcome in $[0, 1]$ and regress it on the auxiliary covariates W_0 by fitting a regression model, the most common choices are beta regression and simplex regression (Kieschnick and McCullough, 2003). Beta regression is a natural choice when one

believes that the outcome follows a beta distribution, while simplex regression is used if the empirical distribution of the sampling probability resembles a bimodal pattern (Zhang et al., 2014; Espinheira and Silva, 2018). We estimate $P(R_1 = 1 \mid W_0, R_{\text{comb}} = 1)$ using the combined data consisting of both the EHR sample and the external probability sample by fitting a generalized linear model. Finally, we estimate $\lambda_1(W_0)$, the probability of being selected into the phase-I EHR sample, via Eq. (3).

3.2.2. *Method 2: indirectly addressing selection bias by subsampling strategy.* Instead of directly estimating $\lambda_1(W_0)$, an alternative approach to address selection bias is to draw an unbiased subsample from the biased phase-I EHR sample, such that the joint distribution (or at least summary statistics) of W_0 in this subsample is the same as that of the target population. We refer to this subsample as the new phase-I sample, which can be treated as a random sample of the target population. A practical approach to obtain such an unbiased subsample from the biased EHR sample is through matching. Specifically, we first simulate a sample of W_0's according to the distribution of W_0 in the target population, then we match the biased EHR sample to the unbiased W_0 data using the nearest neighbor matching method without replacement (Ho et al., 2007; Stuart, 2010). This matching method coincides with the idea of "template matching" proposed by Bennett, Vielma and Zubizarreta (2020).

The new phase-I sample selected by matching is a random sample of the target population. Therefore, the methods proposed by Gilbert, Yu and Rotnitzky (2014) immediately apply. Following our notation and study settings, below we provide the optimal phase-II sampling design under subsampling, which is analogous to Result 3 in Gilbert, Yu and Rotnitzky (2014).

Let $R_1' \in \{0, 1\}$ be the indicator of whether an individual in the target population is selected into the new phase-I sample. Let n_e' ($n_e' \leq n_e$) denote the sample size of the new phase-I sample, then the probability of being selected into the new phase-I sample from the target population is a constant given by $\lambda_1^{\text{alt}} = n_e'/n$. Similar to Theorem 3.1, the optimal phase-II sampling probability under the new phase-I sample is

$$\lambda_2^{\text{alt}}(\overline{W} = \overline{w}) = \min\left\{1, \frac{B - C_0 - n_e' C_1}{n\lambda_1^{\text{alt}}} \frac{\sqrt{Var(Y \mid \overline{W}, R_1' = 1)/C_2(\overline{w})}}{E\left\{\sqrt{C_2(\overline{W})Var(Y \mid \overline{W}, R_1' = 1)}\right\}}\right\}, \tag{4}$$

where $Var(Y \mid \overline{W}, R_1' = 1) = Var(Y \mid \overline{W})$. We refer to the optimal phase-II sampling probability under a new phase-I sample as the alternative two-phase sampling design.

3.3. *Mean outcome estimation.* After the phase-II study sample is recruited and the outcome is measured, we estimate β via the RR estimator (Rotnitzky and Robins, 1995)

$$\widehat{\beta}_{RR} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_{1i}R_{2i}}{\widehat{\lambda}_{1i}\lambda_{2i}}Y_i - \frac{R_{1i}R_{2i} - R_{1i}\lambda_{2i}}{\widehat{\lambda}_{1i}\lambda_{2i}}\widehat{E}(Y \mid \overline{W}_i, R_{1i} = 1) - \frac{R_{1i} - \widehat{\lambda}_{1i}}{\widehat{\lambda}_{1i}}\widehat{E}(Y \mid W_{0i})\right\}.$$

Note that R_1 should be replaced with R_1' if method 2 was used.

A key step is to estimate the outcome models $E(Y \mid W_0)$ and $E(Y \mid \overline{W}, R_1 = 1)$. Recall that by Assumption (1), we have $E(Y \mid W_0) = E\{E(Y \mid \overline{W}, R_2 = 1) \mid W_0, R_1 = 1\}$ and $E(Y \mid \overline{W}, R_1 = 1) = E(Y \mid \overline{W}, R_2 = 1)$. Thus these models can be estimated using outcomes measured in the phase-II sample or from some existing pilot data. We detail the estimation strategies in Section 2.4 of the Supplementary Material (Zhang et al., 2022a). A summary of the overall procedure for sampling design and estimation is presented in Figure 3. We first estimate $Var(Y \mid \overline{W}, R_1 = 1)$ and $\lambda_1(W_0)$ to obtain the optimal phase-II sampling

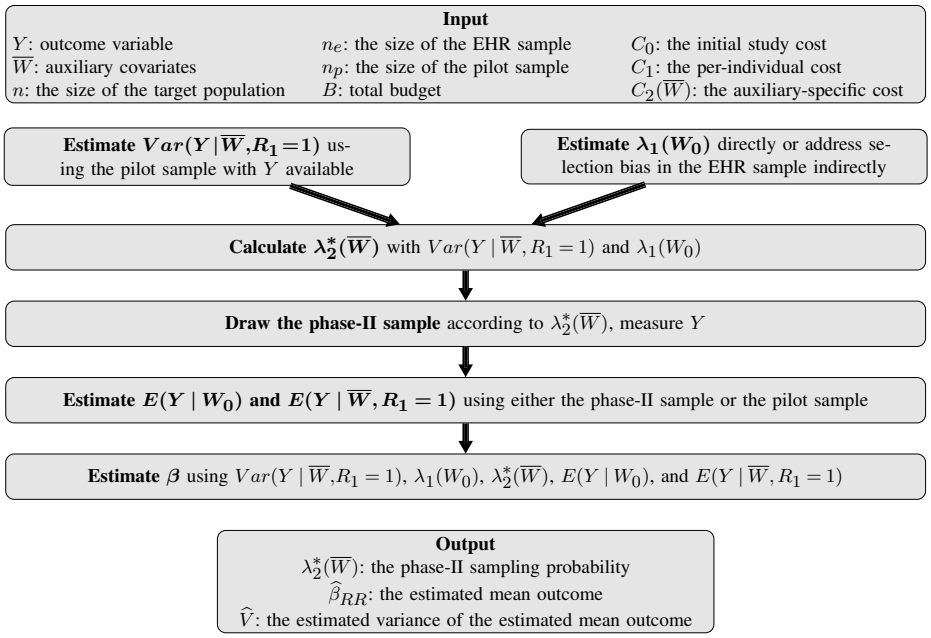

FIG 3. *General Procedure for Efficient Two-Phase Sampling and Mean Estimation.*

design, $\lambda_2^*(\overline{W})$, based on Eq. (2). Then we draw the phase-II sample based on $\lambda_2^*(\overline{W})$ and measure the expensive outcome for each individual in the phase-II sample. Finally, we estimate the parameter based on the proposed RR estimator.

4. Relative efficiency comparing optimal two-phase sampling and random sampling. In this section, we show that our proposed optimal two-phase sampling design is more efficient than random sampling. We define the relative efficiency (RE) comparing the asymptotic variance of the RR estimator obtained using data from optimal phase-II sampling to that of the RR estimator but using data from random phase-II sampling as

$$\text{RE} = \frac{V\{\lambda_2^*(\overline{W})\}}{V(\overline{\lambda}_2)},$$

where $\overline{\lambda}_2 = (B - C_0 - n_e C_1)/[n\lambda_1(W_0)E\{C_2(\overline{W})\}]$ is the random sampling probability derived under the budget constraint. The RE measures the efficiency gain purely due to the proposed optimal auxiliary-dependent sampling design. When $\text{RE} \leq 1$, $V\{\lambda_2^*(\overline{W})\} \leq V(\overline{\lambda}_2)$. The smaller RE is, the more efficiency gain we obtain from the optimal sampling. Further define the proportion of the variation in the outcome explained (PVE) by covariates $\overline{W}$ in the phase-I EHR sample as $\text{PVE} = Var\{E(Y \mid \overline{W}, R_1 = 1)\}/Var(Y)$. The PVE measures how well $\overline{W}$ predicts the outcome Y. Below we present RE under the two methods proposed in Section 3.2 for addressing selection bias in the phase-I sample.

COROLLARY 4.1.

(i) Suppose B and n_e satisfy $n_e < (B - C_0)/C_1$. If selection bias is addressed via direct estimation of selection mechanism (Section 3.2.1), the relative efficiency comparing the proposed optimal sampling to random phase-II sampling is

$$RE = \frac{PVE * Var(Y) + E'\{\lambda_1(W_0)\} + E\left\{\frac{1}{\lambda_2^*(\overline{W})\lambda_1(W_0)} Var(Y \mid \overline{W}, R_1 = 1)\right\}}{PVE * Var(Y) + E'\{\lambda_1(W_0)\} + E\left\{\frac{1}{\overline{\lambda}_2\lambda_1(W_0)} Var(Y \mid \overline{W}, R_1 = 1)\right\}} \leq 1,$$

where $E'\{\lambda_1(W_0)\} = E\left[(1-\lambda_1(W_0))\Big\{Var(Y \mid W_0) - Var(Y \mid \overline{W}, R_1 = 1)\Big\}/\lambda_1(W_0)\right]$.

(ii) If selection bias is addressed indirectly via subsampling (Section 3.2.2), there is no clear pattern of $RE = V\{\lambda_2^{alt}(\overline{W})\}/V(\overline{\lambda}_2)$ *where* $\overline{\lambda}_2$ *is random phase-II sampling from the original phase-I EHR sample. Nevertheless, the relative efficiency comparing the optimal phase-II sampling to random phase-II sampling (both sampling from the new phase-I sample)*

$$RE^{alt} = \frac{V\{\lambda_2^{alt}(\overline{W})\}}{V(\overline{\lambda})} \leq 1$$

when $n_e' < (B - C_0)/C_1$, *where* $\overline{\lambda} = (B - C_0 - n_e' C_1)/[n_e' E\{C_2(\overline{W})\}]$ *is the random phase-II sampling from the new phase-I sample under the budget constraint.*

Corollary 4.1 is proved in Section 2.5 of the Supplementary Material (Zhang et al., 2022a). Corollary 4.1 (i) states that our proposed optimal sampling design improves efficiency compared to random sampling design. Corollary 4.1 (ii) states that if one starts from the new phase-I sample of size n_e' (ignoring the additional $n_e - n_e'$ data points in the original phase-I EHR sample), which is a random sample from the target population, then optimal sampling improves efficiency compared to random sampling. This conclusion about RE^{alt} is analogous to Eq. (11) of Gilbert, Yu and Rotnitzky (2014). We also found that RE and RE^{alt} increase with PVE, which indicates that we obtain less efficiency gain from the optimal sampling design when PVE is large, i.e., the auxiliary covariates $\overline{W}$ explains a major portion of the variation in the outcome in the phase-I EHR sample. Intuitively, if $\overline{W}$ is sufficiently informative in predicting the outcome, then efficiency can be largely achieved by estimation with outcome imputation alone. As such, efficiency gain due to sampling is less obvious when the predictive power of $\overline{W}$ is higher.

Although there is no clear pattern of the relative efficiency comparing $V\{\lambda_2^{\text{alt}}(\overline{W})\}$ and $V(\overline{\lambda}_2)$, we study how it increases or decreases with n_e' in Section 2.5 of the Supplementary Material. We have also provided Table C1 in Section 2.9 of the Supplementary Material to summarize and clarify the difference between sampling methods mentioned in this paper.

5. Extension to a general design framework. Our proposed two-phase sampling design suggests a general auxiliary-covariate-based sampling framework to improve efficiency: for a given estimator which is the solution to an estimating equation that incorporates the selection mechanism in each phase, one derives the optimal phase-II sampling probability by minimizing the asymptotic variance of the estimator under a budget constraint. Based on this idea, in this section, we extend our methods to the estimation of a general parameter of interest, β, defined as the unique solution of an estimating equation $E\{m(Y, \overline{W}; \beta)\} = 0$.

When $(Y, \overline{W})$ is observed on a random sample of the target population, it is straightforward to estimate β by solving $\sum_i m(Y_i, \overline{W}_i; \beta) = 0$. In contrast, the EHR-based two-phase sampling framework constitutes a monotone missing data pattern with observations $O_i = (Y_i R_{2i}, W_{1i} R_{1i}, W_{0i})$, $i = 1, \ldots, n$, as discussed in Section 2.2. In this setting, Tsiatis (2006) presented an augmented inverse probability weighted complete-case estimator for β that solves $E[U(O; \beta)] = 0$, where

(5)

$$U(O;\beta) = \frac{R_1 R_2}{\lambda_1(W_0)\lambda_2(\overline{W})} m(Y, \overline{W}; \beta) - \frac{R_1 R_2 - \lambda_2(\overline{W}) R_1}{\lambda_1(W_0)\lambda_2(\overline{W})} E\{m(Y, \overline{W}; \beta)) \mid \overline{W}, R_1 = 1\}$$
$$- \frac{R_1 - \lambda_1(W_0)}{\lambda_1(W_0)} E\{m(Y, \overline{W}; \beta) \mid W_0\}.$$

One can then derive the optimal sampling probability $\lambda_2(\overline{W})$ based on our proposed framework. We now illustrate this framework with two examples: (1) estimation of causal effects in observational studies or RCTs, and (2) estimation of regression coefficients.

5.1. *Design and estimation for causal inference.* Following the potential outcome framework (Neyman, 1923; Rubin, 1974, 2005), we define a pair of potential outcomes $(Y(1), Y(0))$, representing the outcomes had an individual received treatment or control. Consider a binary treatment A, with $A = 1$ if an individual received treatment and 0 otherwise. Under the standard assumption of consistency, we observe outcome $Y = Y(a)$ if $A = a$, for $a = 0, 1$. Let $\beta_a = E\{Y(a)\}$, $a = 0, 1$, denote the mean potential outcome. The ATE is a contrast between β_1 and β_0 on a user-specified scale, such as $\beta_1 - \beta_0$ or β_1/β_0, depending on the type of outcome and scientific question of interest. Therefore, we focus on deriving the optimal study design for estimation of β_a, $a = 0, 1$.

We will consider optimal study designs for two treatment mechanisms: observational studies and RCTs. In both settings, we will measure the outcome prospectively, with a slight distinction in whether the treatment of interest, A, is readily available in EHR. Specifically, to conduct an observational study, we assume that A is available in the phase-I EHR sample. For each treatment group $A = a$, we aim to select a study cohort and measure the outcomes $Y(a)$ prospectively. To conduct an RCT, we first draw the phase-II study sample, then randomize the recruited study participants to treatment and control and follow up to measure the outcomes of participants.

An important observation is that, in both scenarios, $Y(a)$ is observed only for patients who are selected into the study sample (with $R_2 = 1$) and assigned with treatment a (with $A = a$), for $a = 0, 1$, and $Y(a)$ is missing otherwise. Therefore the composite indicator $\mathbb{I}(R_2 = 1)\mathbb{I}(A = a)$ indicates missingness of the outcome of interest, $Y(a)$. For each $a \in \{0, 1\}$, let

$$Y^\dagger = Y(a), \quad R_{2a} = \mathbb{I}(R_2 = 1)\mathbb{I}(A = a) = R_2\mathbb{I}(A = a), \quad \text{and}$$
$$\lambda_{2a}(\overline{W}) = P(R_{2a} = 1 \mid \overline{W}, R_1 = 1),$$

then there is a correspondence to the basic setting considered in Sections 2-4, in that $Y^\dagger$, R_{2a}, and $\lambda_{2a}(\overline{W})$ can be viewed as Y, R_2, and λ_2, respectively. As such, one can derive an estimator for β_a that incorporates the selection mechanism in each phase and the optimal two-phase sampling probability $\lambda^*_{2a}(\overline{W})$ following the same procedure as Section 3.1.

We first modify the existing assumptions correspondingly.

1†. Missing at random: $(R_2, A) \perp\!\!\!\perp Y^\dagger \mid R_1, \overline{W}$ and $R_1 \perp\!\!\!\perp (Y^\dagger, W_1) \mid W_0$;
2†. Positivity: $\epsilon < \lambda_1(W_0), \lambda_{2a}(\overline{W}) \leq 1$ for some $\epsilon > 0$;
3†. One of the following two sets of models is correctly specified:
(i) Selection probabilities: $\lambda_1(W_0)$ and $\lambda_{2a}(\overline{W})$;
(ii) Outcome regression model: $E(Y^\dagger \mid \overline{W}, R_1 = 1)$ and $f(W_1 \mid W_0, R_1 = 1)$ (thus $E(Y^\dagger \mid W_0, R_1 = 1)$ is also correctly specified).

Now replacing $R_2/\lambda_2(\overline{W})$ with $R_{2a}/\lambda_{2a}(\overline{W})$ in Eq. (1) and noticing that $R_{2a}Y^\dagger = R_{2a}Y$, we have the following estimating function for the mean potential outcome β_a, $a = 0, 1$

$$U_a(O^\dagger; \beta_a) = \frac{R_1 R_{2a}}{\lambda_1(W_0)\lambda_{2a}(\overline{W})}Y - \frac{R_1 R_{2a} - R_1\lambda_{2a}(\overline{W})}{\lambda_1(W_0)\lambda_{2a}(\overline{W})}g(\overline{W}) - \frac{R_1 - \lambda_1(W_0)}{\lambda_1(W_0)}\tilde{g}(W_0) - \beta_a,$$

where $O^\dagger = (Y^\dagger R_{2a}, W_1 R_1, W_0)$, $g(\overline{W}) = E(Y^\dagger \mid \overline{W}, R_1 = 1)$, and $\tilde{g}(W_0) = E(Y^\dagger \mid W_0)$. In fact, $U_a(O^\dagger; \beta_a)$ is a special case of Eq. (5), with $m(Y^\dagger, \overline{W}; \beta_a) = Y^\dagger - \beta_a$.

To derive the optimal design for estimation of β_a, we consider a general setting where the costs for measuring outcomes for patients exposed to different treatments are allowed to be different. Let B_a denote the total budget for estimating β_a, C_{0a} denote the initial study cost for estimating β_a, C_1 denote the per-individual cost in phase-I that scales with size of an EHR sample, and $C_{2a}(\overline{W})$ denote the per-individual cost in phase-II that may depend on auxiliary covariates for patient characteristics. We define phase-I cost of accessing the EHR sample,

$n_{ea}C_1$, as follows for an RCT or observational study: in an RCT, $n_{ea}C_1$ is the phase-I cost distributed to the treatment arm a according to phase-II treatment allocation proportion; in an observational study, $n_{ea}C_1$ is the cost of accessing EHR samples in treatment arm a. By Theorem 3.1, we have the following optimal probability of being selected into the phase-II sample and being assigned with treatment a

$$\lambda_{2a}^*(\overline{W}=\overline{w})=\min\left\{1,\frac{B_a-C_{0a}-n_{ea}C_1}{n\lambda_1(w_0)}\frac{\sqrt{Var(Y^\dagger\mid\overline{W},R_1=1)/C_{2a}(\overline{w})}}{E\left\{\sqrt{C_{2a}(\overline{W})Var(Y^\dagger\mid\overline{W},R_1=1)}\right\}}\right\}.$$

Similar to Section 3.1, $Var(Y^\dagger\mid\overline{W},R_1=1)$ can be estimated using a pilot sample following the procedure detailed in Section 2.7 of the Supplementary Material (Zhang et al., 2022a).

Note that

$$\lambda_{2a}(\overline{W})=P(R_2=1\mid A=a,\overline{W},R_1=1)P(A=a\mid\overline{W},R_1=1) \tag{6}$$

$$=P(A=a\mid R_2=1,\overline{W},R_1=1)P(R_2=1\mid\overline{W},R_1=1). \tag{7}$$

For designing an observational study with treatment A readily available in the phase-I sample, we first estimate $P(A=a\mid\overline{W},R_1=1)$ using the entire EHR data under a pre-specified model, then derive the optimal phase-II sampling probability for each treatment group, i.e., $P(R_2=1\mid A=a,\overline{W},R_1=1)$, based on Eq. (6). For designing an RCT, because treatment A is randomized, we have $P(A=1\mid R_2=1,\overline{W},R_1=1)=c$, where $c\in(0,1)$ is a constant. For a given c, we derive the optimal phase-II sampling probability, i.e., $P(R_2=1\mid\overline{W},R_1=1)$ based on Eq. (7). We present details on the derivation in Section 2.6 of the Supplementary Material (Zhang et al., 2022a).

Once the phase-II sample is drawn and $Y(a)$ is measured, one can estimate β_a by solving $n^{-1}\sum_{i=1}^n U_a(O_i^\dagger;\beta_a)=0$, with similar estimation procedure as Section 3.3. A key step is to estimate $g(\overline{W})$ and $\tilde{g}(W_0)$. Under Assumption $1^\dagger$ we have $g(\overline{W})=E(Y\mid\overline{W},R_1=1,A=a)$ and $\tilde{g}(W_0)=E\{g(\overline{W})\mid W_0,R_1=1\}$, hence $g(\overline{W})$ and $\tilde{g}(W_0)$ can be estimated using outcomes measured in phase-II sample or from pilot data. We detail the estimation strategy in Section 2.7 of the Supplementary Material, and present the overall algorithm in Section 2.8 of the Supplementary Material (Zhang et al., 2022a).

We further note that the above estimator of β_a is doubly robust in the sense that it is consistent if either the selection probability models or the outcome regression models are correctly specified, i.e., Assumption $3^\dagger$ holds, which we prove in Section 2.6 of the Supplementary Material (Zhang et al., 2022a).

5.2. *Design and estimation for regression models.* Another example is the coefficient β in a regression model, in which case $m(Y,\overline{W};\beta)=d(\overline{W})\{Y-g(\overline{W};\beta)\}$, where $g(\overline{W};\beta)=E(Y\mid\overline{W};\beta)$ is a user-specified outcome regression model indexed by coefficient β, and $d(\overline{W})$ is a vector of functions of $\overline{W}$ of the same dimension as β. Here we briefly discuss extension to the estimation of β.

Let $\epsilon(\beta)=Y-g(\overline{W};\beta)$. Under Assumptions 1-3, Tsiatis (2006) proposed an augmented inverse probability weighted estimator for β that solves $E[U(O;\beta)]=0$, where

$$U(O;\beta)=\frac{R_1R_2}{\lambda_1(W_0)\lambda_2(\overline{W})}d(\overline{W})\epsilon(\beta)-\frac{R_1R_2-\lambda_2(\overline{W})R_1}{\lambda_1(W_0)\lambda_2(\overline{W})}E\{d(\overline{W})\epsilon(\beta)\mid\overline{W},R_1=1\}$$
$$-\frac{R_1-\lambda_1(W_0)}{\lambda_1(W_0)}E\{d(\overline{W})\epsilon(\beta)\mid W_0\}.$$

The asymptotic variance for the estimator is given by $V(\lambda_2) = \tau^{-1}(O)Var(U)\{\tau^{-1}(O)\}^\top$, where $\tau(O) = E\{\partial u(O;\beta)/\partial\beta^\top\}$ and $u(O;\beta) = R_1R_2d(\overline{W})\epsilon(\beta)/\lambda_1(W_0)\lambda_2(\overline{W})$. Thus one can derive the optimal phase-II sampling probability $\lambda_2^*(\overline{W})$ by minimizing $V(\lambda_2)$ under a budget constraint.

6. Simulation. We conduct simulation studies to assess the performance of the optimal two-phase sampling design and the RR estimator in finite samples. We first generate a target population of size $n = 10,000$. We then emulate the process of selection into the phase-I EHR sample from the target population according to a pre-specified phase-I sampling probability, and the resulting EHR sample size n_e is approximately $5,000$. We also randomly draw a pilot sample of size $n_p = 200$ from the EHR sample. The data generating mechanism is as follows.

- $W_0 \sim N(0.05, 2)$: patient characteristics that impact selection into the phase-I EHR sample
- $R_1 \mid W_0 \sim$ Bernoulli$\{p = \lambda_1(W_0)\}$: indicator for selection into the phase-I sample with
  1. $\lambda_1(W_0) = (1 + e^{-W_0})^{-1}$: modest selection bias setting
  2. $\lambda_1(W_0) = 0.9\mathbb{I}(W_0 > 0.08) + 0.1\mathbb{I}(W_0 \leq 0.08)$: extreme selection bias
- $W_1 \sim N(0.05, 2)$: additional patient characteristics observed from the phase-I sample
- $Y \mid \overline{W} \sim N\{E(Y \mid \overline{W};\boldsymbol{\alpha}), Var(Y \mid \overline{W};\boldsymbol{\gamma})\}$: the observed outcome, where $E(Y \mid \overline{W};\boldsymbol{\alpha}) = \alpha_0 + \alpha_1W_0 + \alpha_2W_1$ with $\boldsymbol{\alpha} = (0.1, 3, 0.01)$, and $Var(Y \mid \overline{W};\boldsymbol{\gamma}) = \exp(\gamma_{00} + \gamma_0W_0 + \gamma_1W_0^2 + \gamma_2W_1 + \gamma_3W_1^2)$. We set $\gamma_{00} = -1.5$, $\gamma_1 = 0.2$, $\gamma_2 = \gamma_3 = 0.01$, and we consider three values of γ_0: 0.97, 0.82, and -0.64, which correspond to the scenarios where the PVE is low (0.2), moderate (0.5), and high (0.8), respectively.

The model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are estimated via maximum likelihood estimation as detailed in Section 2.4 of the Supplementary Material (Zhang et al., 2022a). We further define the total budget and various costs as follows: the total budget is $B = 100,000$, the initial study cost is $C_0 = 50,450$, the per-individual cost for obtaining the EHR data is $C_1 = 0.01$, and the auxiliary-specific per-individual cost for measuring the outcome of interest is $C_2(\overline{W}) = 100$.

We compare the performance of four approaches listed in Table 1 which differ in sampling design and estimation method.

TABLE 1

Four approaches for sampling design and parameter estimation. Approach 1: (naive) measure outcomes from a random sample of the phase-I sample and estimate β using the sample mean. Approach 2: (Random sampling and the RR estimator) measure outcomes from a random sample of the phase-I sample and estimate β using the RR estimator. Approach 3: (Optimal sampling and the RR estimator) model phase-I selection bias $\lambda_1(W_0)$, then measure outcomes based on our proposed optimal two-phase sampling design, and estimate β using the RR estimator.

Approach	Legend	Method for		Estimation	Model Specification for	
		Phase-I Selection Bias	Phase-II Sampling Design		$E[Y \mid W_0]$ and $E[Y \mid \overline{W}, R_1 = 1]$	$Var(Y \mid \overline{W}, R_1 = 1)$
1	○	Modeling†	Random	Sample mean	N/A	N/A
2	△	Modeling	Random	RR estimator	Correct	Correct
3a	+	Modeling	Optimal††	RR estimator	Correct	Correct
3b	×	Modeling	Optimal	RR estimator	Correct	Misspecified
3c	◇	Modeling	Optimal	RR estimator	Misspecified	Correct
3d	▽	Modeling	Optimal	RR estimator	True model	True model

†Method 1 for handling selection bias in the phase-I EHR sample, detailed in Section 3.2.1. In simulation studies, we use the true $\lambda_1(W_0)$, because performance of the estimation method has been thoroughly evaluated by Beesley and Mukherjee (2020) and will be further illustrated in our application study in Section 7.
‡Method 2 for handeling selection bias in the phase-I EHR sample, detailed in Section 3.2.2.
††Approach 3 handles phase-I selection bias by modeling $\lambda_1(W_0)$ with corresponding optimal phase-II sampling probability presented in Eq. (2).

We investigate the performance of the above approaches in terms of relative efficiency computed from $50,000$ Monte Carlo replications. Relative efficiency is measured by the empirical variance of the estimates from each approach standardized by the empirical variance

of the estimates from Approach 2. Lower relative efficiency value indicates lower variance of $\widehat{\beta}_{RR}$ and hence a more efficient sampling approach relative to Approach 2.

In Figure 4, we present the relative efficiency of Approaches 1-3d compared to Approach 2 under modest and extreme selection bias, as well as low, moderate, and high PVE. Across all PVEs and in both panels, we can see that the RE of Approach 1 compared to Approach 2 is larger than one, which indicates that the RR estimator that utilizes auxiliary information is more efficient than a simple average of the measured outcomes from phase-II. Comparing Approach 3 and Approach 2, the RE is generally smaller than one, which implies that our proposed optimal two-phase sampling design improves efficiency over random sampling. Approach 3b with $Var(Y \mid \overline{W}, R_1 = 1)$ misspecified is the least efficient among all scenarios of Approach 3, which is expected because the optimal sampling design depends on the estimation of $Var(Y \mid \overline{W}, R_1 = 1)$. Misspecification of the variance model can have a notable impact on efficiency, while misspecification of the mean models (Approach 3b) does not have a substantial impact.

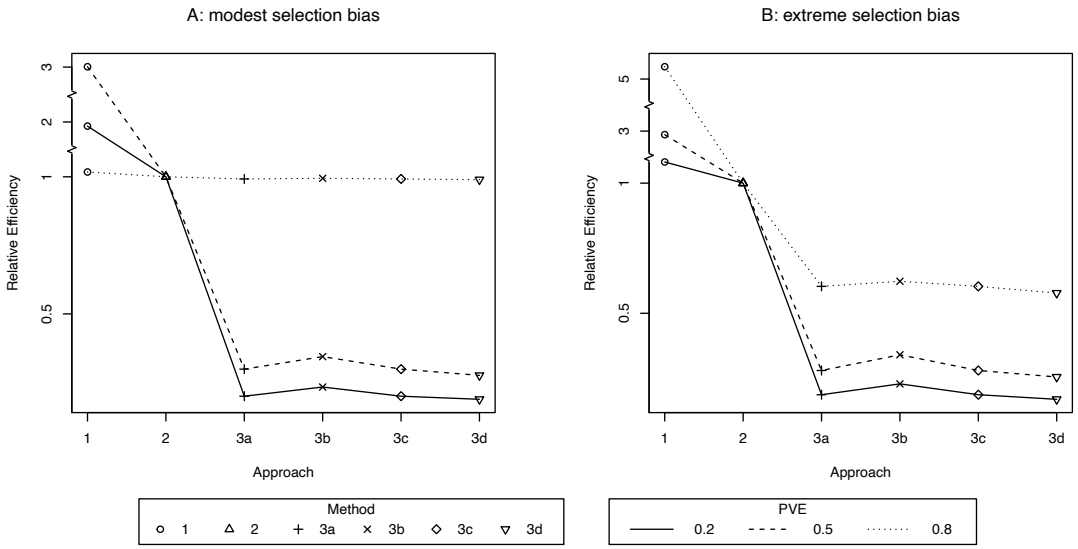

FIG 4. *The results of simulation studies. The relative efficiency of Approach 1 and Approach 3a-3c compared to Approach 2, under modest (panel A) and extreme (panel B) selection bias, as well as low (*0.2*), moderate (*0.5*), and high (*0.8*) PVE is presented. The relative efficiency is measured by the Monte Carlo variance of the estimates from each approach divided by that from Approach 2. Lower relative efficiency value indicates smaller estimation variance relative to Approach 2 and hence a more efficient approach. We study the following approaches: (1) naive; (2) random sampling and the RR estimator; (3) optimal sampling and the RR estimator under (3a) correctly specified mean and variance models; (3b) misspecified variance model; (3c) misspecified mean model; and (3d) true models. PVE stands for the proportion of the variation in the outcome explained, which is defined in Section 4.*

We further investigate the role of PVE which measures the predictive accuracy of the auxiliary covariates $\overline{W}$. In both panels A and B, the RE of Approach 3 compared to Approach 2 increases with PVE, which is consistent with the conclusion of Corollary 4.1. We conduct sensitivity analysis under a relatively smaller pilot sample size ($n_p = 50$). The results are presented in Figure B4 of Section 3 of the Supplementary Material (Zhang et al., 2022a). We observe similar efficiency gain from the optimal two-phase sampling approaches particularly under modest selection bias, while the small pilot data may not be sufficient under an

extremely biased phase-I sample. We also show the relative efficiency comparing the alternative phase-II sampling design to random sampling under different settings in Section 3 of the Supplementary Material (Zhang et al., 2022a).

7. Estimating the prevalence of hypertension among US adults using MGI data. We estimate the prevalence of hypertension in the US adult population using EHR data from the MGI. A flowchart of the two-phase sampling procedure and data summaries is presented in Figure 1. We take the 2017-2018 NHANES data on $6,724$ individuals aged 18 and older as our external probability sample of the target population to estimate $\lambda_1(W_0)$ following methods in Section 3.2.1. We choose hypertension as our outcome of interest because it is readily measured in MGI and NHANES data, which allows us to validate our results. Using inverse probability of sampling weighting, we estimated from NHANES data that the benchmark prevalence of hypertension in the target population is 39.19%.

Let $Y \in \{0, 1\}$ denote the indicator of being diagnosed with hypertension, then our parameter of interest $\beta = P(Y = 1)$, where the expectation is taken with respect to the distribution of Y in the US adult population. Hypertension status is best determined by chart review, which is time-consuming and resource-intensive. Due to the limitation of resources, we use traditional rule-based phenotyping method instead, which is commonly used to identify hypertension (Chang et al., 2016; Fox et al., 2014). We classify hypertension status from MGI data using the following criteria: Y is unknown for patients who did not receive a hypertension diagnostic procedure (defined by the Current Procedural Terminology (CPT) codes listed in Table C2 of Section 4.1 of the Supplementary Material (Zhang et al., 2022a)); for those who received at least one hypertension diagnostic procedure, we define $Y = 1$ if there is the presence of at least one hypertension diagnosis code in the patient's record (defined by the International Classification of Diseases (ICD) codes listed in Table C2 of Section 4.1 of the Supplementary Material (Zhang et al., 2022a)), and $Y = 0$ otherwise. We take age, sex, and race as W_0 and take smoking status and body mass index (BMI) as W_1, which are risk factors associated with hypertension (Brown et al., 2000; Pinto, 2007). We categorize BMI as follows: underweight if BMI < 18.5, normal weight if BMI $\in [18.5, 25)$, overweight if BMI $\in [25, 30)$, and obese if BMI ≥ 30. We assume the target population size is $n = 330,000,000$. We consider two scenarios where the total budget B is equal to either $100,000$ or $1,000,000$. We assume various study costs as follows: the initial study cost is $C_0 = 10,000$, the per-individual cost for obtaining the EHR sample is $C_1 = 0.01$, and the auxiliary-specific per-individual cost for measuring the outcome is $C_2(\overline{W}) = 100 + 5 \times \text{age} + 5 \times \text{gender} + 5 \times \text{race} + 5 \times \text{BMI} + 5 \times \text{smoking}$.

We define the phase-I EHR sample under the following exclusion criteria. We exclude MGI individuals with missing age, gender, race, smoking status, BMI status, and individuals under the age of 18. We further exclude individuals who did not receive any diagnostic procedure for hypertension, such that any pilot sample randomly selected from the EHR sample have outcome data available for illustration purpose. After data cleaning, the phase-I EHR sample size is $n_e = 80,934$. The majority of individuals in the EHR sample are over 60 (47.51%), female (53.92%), white (84.49%), never-smokers (54.72%), and obese (42.93%). In addition, hypertension is more prevalent among individuals who are over 60 (69.25%), male (54.96%), black (57.58%), former smokers (60.18%), and obese (64.28%). As a consequence, the raw prevalence of hypertension is 49.19% in MGI, which is much higher than the benchmark value and indicates selection bias in the EHR sample.

We apply Approaches 1-3 investigated in the simulation study. We assume that $Var(Y \mid \overline{W}; \boldsymbol{\gamma}) = \exp(\gamma_{00} + \gamma_0 \text{age} + \gamma_1 \text{sex} + \gamma_2 \text{race} + \gamma_3 \text{BMI} + \gamma_4 \text{smoking})$, $E(Y \mid W_0; \boldsymbol{\delta}) = \delta_{00} + \delta_0 \text{age} + \delta_1 \text{sex} + \delta_2 \text{race}$, and $E(Y \mid \overline{W}; \boldsymbol{\alpha}) = \alpha_{00} + \alpha_0 \text{age} + \alpha_1 \text{sex} + \alpha_2 \text{race} + \alpha_3 \text{BMI} + \alpha_4 \text{smoking}$. We randomly draw 100 individuals from the EHR sample as the pilot data to

estimate parameters in the variance model, while we use the phase-II study sample to estimate parameters in the mean models. We address selection bias in EHR data by modeling (detailed in Section 3.2.1). We choose to fit a beta regression model which is supported by a goodness-of-fit test and the observation that the empirical distribution of the sample probability in the external probability sample is unimodal. We obtain estimates of β from Approaches 3. We use the sample variance of $50,000$ bootstrapped estimates to calculate the relative efficiencies of Approach 3 versus Approach 1 and Approach 3 versus Approach 2, denoted as $\text{RE}_{3\text{vs}1}$ and $\text{RE}_{3\text{vs}2}$, respectively. We clarify that we aim to apply our optimal sampling to improve efficiency of the RR estimator that accounts for selection bias, and the sampling design itself does not aim to reduce selection bias by creating a study sample with similar characteristics as the target population. Thus, the estimated hypertension prevalence is mainly used to compare the relative efficiency of different sampling approaches rather than to demonstrate reduction of selection bias.

Figure 5 presents smoothed curves for the relationship between the designed $\lambda_2^*(\overline{W})$ from Approach 3 and two selected auxiliary covariates, which are age and BMI. We can see that the optimal sampling probability decreases with both age and BMI, which implies that the proposed method would oversample younger individuals with lower BMI, who are less likely to have hypertension (Brown et al., 2000). This is expected because, as discussed previously, the prevalence of hypertension in the EHR sample (49.19%) is higher than that of the target population (39.19%). Our proposed sampling design is able to distinguish the compositions of patient characteristics between the EHR sample and the target population. As a result, individuals who are less likely to have hypertension are oversampled by our proposed sampling design, which will lead to a final estimate that is lower than the raw prevalence in the EHR sample. We also observe that as the total budget B increases, $\lambda_2^*(\overline{W})$ uniformly increases. This is also expected because, as budget increases, we can afford a larger study sample and thus individuals in the EHR sample are generally more likely to be sampled.

We also present the estimated hypertension prevalence and relative efficiency comparing different approaches in the legend of Figure 5. By Approach 3, when $B = 10^5$, the size of the selected study sample is approximately $n_s = 100$ and the estimated prevalence is 38.32% (95% CI: $28.90\%, 47.75\%$); when $B = 10^6$, the study sample size is approximately $n_s = 1,000$ and the estimated prevalence is 39.47% (95% CI: $36.61\%, 42.32\%$). Compared to the raw estimate from MGI data, these estimates are closer to the benchmark value. The estimated $\text{RE}_{3\text{vs}1}$'s and $\text{RE}_{3\text{vs}2}$'s are all smaller than one, which demonstrates the efficiency gain from both the optimal sampling design and from incorporating auxiliary information in estimation.

We consider the US adult population as the target population in the study above, but one may argue that it is more relevant to generalize to the Michigan adult population. Therefore, we perform a sensitivity analysis (details in Section 4.2 of the Supplementary Material (Zhang et al., 2022a)) to estimate the prevalence of hypertension among Michigan adults to further validate the efficiency gain of the proposed two-phase sampling framework. Compared to the raw estimate from MGI data subject to selection bias, the estimate obtained from Approach 3 is more precise with efficiency gain from our derived optimal sampling design compared to random sampling.

8. Discussion. Electronic health records data open up opportunities for cost-effective patient recruitment (McCord and Hemkens, 2019). In this paper, motivated by the observation that recruiting patients from EHR sample constitutes a two-phase sampling framework, we derive the optimal sampling design to minimize the asymptotic variance of an estimator of the mean or mean difference of an outcome of interest that incorporates the selection mechanism in each phase. We extend the two-phase sampling framework proposed by Gilbert, Yu and Rotnitzky (2014) to further account for potential selection bias in EHR data. We

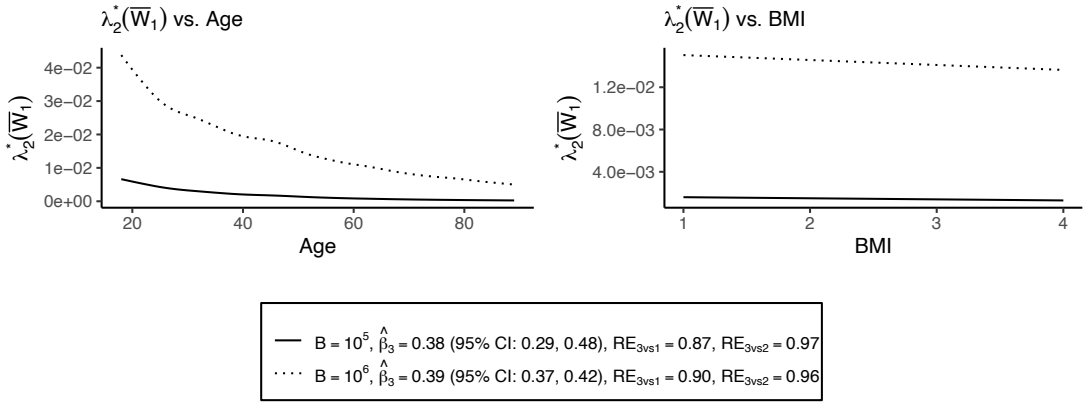

FIG 5. *Results of estimating hypertension prevalence in the US adult population using EHR data from the MGI: the relationship between the optimal second phase sampling probability $\lambda_2^*(\overline{W})$ and auxiliary covariates (age and BMI), as well as the relative efficiency comparing different approaches. This figure shows that $\lambda_2(\overline{W})$ tends to decrease with age and BMI and increase with budget. In the legend, $\widehat{\beta}_3$ is the estimate of β from Approach 3, RE_{3vs1} is the relative efficiency of Approach 3 versus Approach 1, and RE_{3vs2} is the relative efficiency of Approach 3 versus Approach 2.*

highlight that our proposed method is efficient in design and robust in estimation. First, our proposed two-phase sampling design is more efficient than random sampling, which is shown in asymptotic theory, through finite sample simulation studies, and with an application study to real-world EHR data. Second, the RR estimator is doubly robust in the sense that it is consistent under correct specification of either the phase-I selection mechanism or the outcome models. In addition, by incorporating auxiliary information in all phases, the RR estimator is generally more efficient than a simple weighted average of the outcome. We also extend the proposed method to a general two-stage sampling design framework for estimation of a general estimand, such as the ATE and regression coefficients.

There is a rich literature on optimal two-phase sampling and an evolving literature on patient recruitment using EHR (Levis et al., 2022; Barrett et al., 2020; Gilbert, Yu and Rotnitzky, 2014). Barrett et al. (2020) proposed to recruit using EHR such that the distribution of covariate in the study sample is approximately uniformly distributed. Levis et al. (2022) considered the scenario where outcomes in the EHR sample are potentially missing not at random and proposed a double sampling design to further collect outcomes. They established identification and derived nonparametric efficient estimators. They ultimately extended the framework to arbitrary coarsening mechanisms, which includes the two-phase sampling framework of Gilbert, Yu and Rotnitzky (2014) as a special case. It is noteworthy that Gilbert, Yu and Rotnitzky (2014) proposed an optimal three-phase sampling framework in their appendix, which assumed a random phase-I sample and proposed an optimal sampling design for the second and third phases. Our setting differs from that of Gilbert, Yu and Rotnitzky (2014) in that investigators have no control over the selection mechanism of the EHR sample, thus the optimal three-phase sampling framework cannot be directly applied to our setting. We make the following contributions to the literature. First, to our limited knowledge, existing literature did not address the impact of selection bias in EHR on study design. Our work, built upon Gilbert, Yu and Rotnitzky (2014), provides sampling and estimation approaches accounting for selection bias commonly seen in EHR data. Second, we extend the current literature to efficient sampling for the estimation of a general parameter of interest, which covers a wide range of problems such as coefficients in regression models of many kinds,

and the average treatment effect in causal inference. Our proposed method may also be applied to guide the process of selecting a chart review sample to obtain gold standard labels for semi-supervised learning and building phenotyping algorithms (Beaulieu-Jones et al., 2016; Zhang et al., 2022b).

Our proposal has the following limitations. First, EHR systems are designed and optimized for clinical and billing purposes rather than research. As a result, the data collected are subject to a range of issues including missing data, measurement, and classification error, and confounding, which can undermine the validity of any EHR-based research. Second, in practice, modeling the propensity for an individual to seek care is likely challenging and may depend on covariates that are not measured, such as time-varying biomarkers, symptoms, and prior outcomes. The validity of our proposed methods is potentially limited by the amount of prior knowledge on W_0. Third, our positivity assumption for the probability of being selected into the phase-I EHR sample may not be met in practice and should be carefully evaluated. In particular, generalizability may no longer hold when patients with certain characteristic have zero probability of being selected in the EHR sample and such characteristic modifies the outcome model. Lastly, although using pilot data to estimate design parameters is a common approach in two-phase sampling (Gilbert, Yu and Rotnitzky, 2014; McIsaac and Cook, 2015), caution should be taken regarding issues such as the potential selection bias within a pilot study, and the need to further account for uncertainty in estimation based on pilot data when making inference on the parameter of interest. Methods to address these limitations warrant future research.

In summary, we believe our proposed sampling and estimation procedure contributes to the two-phase sampling literature and may shed light on data-driven patient recruitment leveraging large-scale EHR data to facilitate biomedical research.

Acknowledgments. We thank the Michigan Genomics Initiative participants, Precision Health at the University of Michigan, and the University of Michigan Medical School Data Office for Clinical and Translational Research for providing data storage, management, processing, and distribution services. We thank the Advanced Research Computing Technology Services at the University of Michigan for providing data storage and computing resources. The study protocols were reviewed and determined exempt by the University of Michigan Medical School Institutional Review Board (IRB ID HUM00177982).

SUPPLEMENTARY MATERIAL

Supplement to "Patient recruitment using electronic health records under selection bias: a two-phase sampling framework"
We provide additional material to support the results in this paper. Section 1 reviews the literature on two-phase sampling design; Section 2.1 provides a derivation of the total expected cost; Section 2.2 includes a proof of the double robustness of the RR estimator; Section 2.3 includes a proof of Theorem 3.1; Section 2.4 details an estimation procedure for the mean outcome under two-phase sampling; Section 2.5 presents a proof of Corollary 4.1; Section 2.6 includes a proof of the double robustness of the RR estimator for ATE; Section 2.7 details estimation strategies for the ATE under two-phase sampling; Section 2.8 details the overall algorithm for design and estimation of the ATE; Section 2.9 presents a table summarizing different methods mentioned in this paper. Section 3 includes additional results of the simulation study; Section 4.1 lists medical codes used for hypertension phenotyping in the application study; Section 4.2 provides a sensitivity analysis for the application study.

Source code to "Patient recruitment using electronic health records under selection bias: a two-phase sampling framework"
We provide R code for implementing the simulation studies and the application study at `https://github.com/guanghaozhang/TwoPhaseSampling`

REFERENCES

BARRETT, J. E., CAKIROGLU, A., BUNCE, C., SHAH, A. and DENAXAS, S. (2020). Selective recruitment designs for improving observational studies using electronic health records. *Statistics in Medicine* **39** 2556–2567.

BEAULIEU-JONES, B. K., GREENE, C. S. et al. (2016). Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of biomedical informatics* **64** 168–178.

BEESLEY, L. J. and MUKHERJEE, B. (2020). Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*.

BEESLEY, L. J., SALVATORE, M., FRITSCHE, L. G., PANDIT, A., RAO, A., BRUMMETT, C. et al. (2020). The emerging landscape of health research based on biobanks linked to electronic health records: existing resources, statistical challenges, and potential opportunities. *Statistics in Medicine* **39** 773–800.

BENNETT, M., VIELMA, J. P. and ZUBIZARRETA, J. R. (2020). Building representative matched samples with multi-valued treatments in large observational studies. *Journal of computational and graphical statistics* **29** 744–757.

BOWER, J. K., BOLLINGER, C. E., FORAKER, R. E., HOOD, D. B., SHOBEN, A. B. and LAI, A. M. (2017). Active use of electronic health records (EHRs) and personal health records (PHRs) for epidemiologic research: sample representativeness and nonresponse bias in a study of women during pregnancy. *eGEMs: The Journal of Electronic Health Data and Methods* **5** 1263.

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge University Press.

BROWN, C. D., HIGGINS, M., DONATO, K. A., ROHDE, F. C., GARRISON, R., OBARZANEK, E. et al. (2000). Body mass index and the prevalence of hypertension and dyslipidemia. *Obesity Research* **8** 605–619.

CHANG, W.-T., WENG, S.-F., HSU, C.-H., SHIH, J.-Y., WANG, J.-J., WU, C.-Y. and CHEN, Z.-C. (2016). Prognostic factors in patients with pulmonary hypertension—a nationwide cohort study. *Journal of the American Heart Association* **5** e003579.

COWIE, M. R., BLOMSTER, J. I., CURTIS, L. H., DUCLAUX, S., FORD, I., FRITZ, F. et al. (2017). Electronic health records to facilitate clinical research. *Clinical Research in Cardiology* **106** 1–9.

EFFOE, V. S., KATULA, J. A., KIRK, J. K., PEDLEY, C. F., BOLLHALTER, L. Y., BROWN, W. M. et al. (2016). The use of electronic medical records for recruitment in clinical trials: findings from the Lifestyle Intervention for Treatment of Diabetes trial. *Trials* **17** 496.

ELLIOT, M. R. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*.

ESPINHEIRA, P. and SILVA, A. D. O. (2018). Nonlinear Simplex Regression Models. *arXiv preprint arXiv:1805.10843*.

FOX, B. D., AZOULAY, L., DELL'ANIELLO, S., LANGLEBEN, D., LAPI, F., BENISTY, J. and SUISSA, S. (2014). The use of antidepressants and the risk of idiopathic pulmonary arterial hypertension. *Canadian Journal of Cardiology* **30** 1633–1639.

GILBERT, P. B., YU, X. and ROTNITZKY, A. (2014). Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Statistics in Medicine* **33** 901–917.

GOLDSTEIN, B. A., BHAVSAR, N. A., PHELAN, M. and PENCINA, M. J. (2016). Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology* **184** 847–855.

HANEUSE, S. and DANIELS, M. (2016). A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs: The Journal of Electronic Health Data and Methods* **4** 16.

HÄYRINEN, K., SARANTO, K. and NYKÄNEN, P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International Journal of Medical Informatics* **77** 291–304.

HEMKENS, L. G., CONTOPOULOS-IOANNIDIS, D. G. and IOANNIDIS, J. P. (2016). Routinely collected data and comparative effectiveness evidence: promises and limitations. *The Canadian Medical Association Journal* **188** E158–E164.

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15** 199–236.

JOYCE, E., WANG, S., MOTAMED, M., KIDWELL, K. M. and HENRY, N. L. (2021). Associations between preexisting nociplastic pain and early discontinuation of aromatase inhibitor therapy in breast cancer. *Journal of Clinical Oncology* **39** 12068-12068.

KIESCHNICK, R. and MCCULLOUGH, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling* **3** 193–213.

LEVIS, A. W., MUKHERJEE, R., WANG, R. and HANEUSE, S. (2022). Double sampling and semiparametric methods for informatively missing data. *arXiv preprint arXiv:2204.02432*.

MCCORD, K. A. and HEMKENS, L. G. (2019). Using electronic health records for clinical trials: Where do we stand and where can we go? *The Canadian Medical Association Journal* **191** E128–E133.

MCISAAC, M. A. and COOK, R. J. (2015). Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. *Statistics in Medicine* **34** 2899–2912.

NEYMAN, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences* **10** 1–51.

PHELAN, M., BHAVSAR, N. A. and GOLDSTEIN, B. A. (2017). Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. *eGEMs: The Journal of Electronic Health Data and Methods* **5** 22.

PINTO, E. (2007). Blood pressure and ageing. *Postgraduate Medical Journal* **83** 109–114.

ROTNITZKY, A. and ROBINS, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82** 805–820.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66** 688.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100** 322–331.

SAHAY, B., NGUYEN, C. Q. and YAMAMOTO, J. K. (2017). Conserved HIV epitopes for an effective HIV vaccine. *Journal of Clinical & Cellular Immunology* **8**.

SCHREIWEIS, B., TRINCZEK, B., KÖPCKE, F., LEUSCH, T., MAJEED, R. W., WENK, J., BERGH, B., OHMANN, C., RÖHRIG, R., DUGAS, M. and PROKOSCH, H.-U. (2014). Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. *International Journal of Medical Informatics* **83** 860–868.

SHI, X., PAN, Z. and MIAO, W. (2023). Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics* **15** e1581.

SHORTREED, S. M., COOK, A. J., COLEY, R. Y., BOBB, J. F. and NELSON, J. C. (2019). Challenges and opportunities for using big health care data to advance medical science and public health. *American Journal of Epidemiology* **188** 851–861.

STUART, E. A. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* **25** 1.

THADANI, S. R., WENG, C., BIGGER, J. T., ENNEVER, J. F. and WAJNGURT, D. (2009). Electronic screening improves efficiency in clinical trial recruitment. *Journal of the American Medical Informatics Association* **16** 869–873.

TRIPEPI, G., JAGER, K. J., DEKKER, F. W. and ZOCCALI, C. (2010). Selection bias and information bias in clinical research. *Nephron Clinical Practice* **115** c94–c99.

TSIATIS, A. A. (2006). *Improving efficiency and double robustness with coarsened data* In *Semiparametric theory and missing data* 10, 248-251. Springer.

WU, H., TOTI, G., MORLEY, K. I., IBRAHIM, Z., FOLARIN, A., KARTOGLU, I., JACKSON, R., AGRAWAL, A., STRINGER, C., GALE, D. et al. (2017). SemEHR: surfacing semantic data from clinical notes in electronic health records for tailored care, trial recruitment, and clinical research. *The Lancet* **390** S97.

WU, K.-H. H., HORNSBY, W. E., KLUNDER, B., KRAUSE, A., DRISCOLL, A., KULKA, J., BICKETT-HICKOK, R., FELLOWS, A., GRAHAM, S., KALEBA, E. O. et al. (2021). Exposure and risk factors for COVID-19 and the impact of staying home on Michigan residents. *PLoS ONE* **16** e0246447.

ZHANG, P., QIU, Z., PENG, Z. and ZENGUO, Q. (2014). Regression analysis of proportional data using simplex distribution. *Science China Mathematics (Chinese Version)* **44** 89–104.

ZHANG, G., BEESLEY, L. J., MUKHERJEE, B. and SHI, X. (2022a). Supplement to "Patient recruitment using electronic health records under selection bias: a two-phase sampling framework".

ZHANG, Y., LIU, M., NEYKOV, M. and CAI, T. (2022b). Prior Adaptive Semi-supervised Learning with Application to EHR Phenotyping. *Journal of Machine Learning Research* **23** 1–25.

ZOLLA-PAZNER, S. (2004). Identifying epitopes of HIV-1 that induce protective antibodies. *Nature Reviews Immunology* **4** 199–210.