Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Optimal Surrogate-Assisted Sampling for Cost-Efficient Validation of Electronic Health Record Outcomes

Arielle Marks-Anglin[1] | Jianmin Chen[1] | Chongliang Luo[2] | Rebecca Hubbard[3] | Yong Chen[1]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA | [2]Division of Public Health Sciences, Washington University School of Medicine, St Louis, MO, USA | [3]Department of Biostatistics, Brown University School of Public Health, Providence, RI, USA

Correspondence: Yong Chen (ychen123@upenn.edu)

Received: 3 October 2023 | **Revised:** 14 March 2025 | **Accepted:** 2 April 2025

Funding: This work was supported by the National Institutes of Health (Grant Nos. 1R01AG077820, 1R01LM014344, R01AG073435, R01AI130460, R01CA09377, R01CA120562, R01LM012607, R01LM013519, R21AI167418, R21CA143242, R21CA227613, R56AG069880, R56AG074604, RF1G077820, U01CA063731, U01TR003709), and the Patient-Centered Outcomes Research Institute (Grant Nos. ME-2018C3-14899, ME-2019C3-18315).

Keywords: chart review | informative sampling | phenotyping error

ABSTRACT

Electronic Health Record (EHR) databases are an increasingly valuable resource for observational studies. However, misclassification of EHR-derived outcomes due to imperfect phenotyping leads to bias, inflated type I error, and reduced power in risk-factor association studies. On the other hand, manual chart review to validate outcomes is both cost-prohibitive and time-consuming, and a randomly selected validation sample may not yield sufficient cases to support precise model estimation when the disease is rare. Sampling procedures have been developed for maximizing computational and statistical efficiency in settings where the true disease status is known. However, less work has been done in measurement constrained settings, particularly when an informative surrogate outcome is available. Motivated by this gap, we propose an Optimal Subsampling strategy with Surrogate-Assisted Two-step procedure (OSSAT) to guide cost-effective chart review in measurement constrained settings. The sampling weight in OSSAT leverages information contained in the potentially misclassified phenotype and covariates to prioritize observations most informative for the model of interest. We compare our proposed weight with existing approaches through simulations under various covariate distributions, differential misclassification rates and degrees of surrogate accuracy. We then apply our proposed weighting schemes to a study of risk factors for second breast cancer events using a real EHR data set.

1 | Introduction

Electronic health record (EHR) data are increasingly utilized for clinical research due to the tremendous volume of patient data available and the extensive health information contained in them [1, 2], enabling novel investigations and discoveries. The advantages of using EHR data include the ability to leverage information not routinely collected in prospective trials, conduct studies involving rare conditions (e.g., pediatric chronic conditions [3]), evaluate treatment effects in diverse, non-trial populations, identify new indications for drug repurposing [4], and predict adverse events for drug usage [5]. In order to utilize EHRs in clinical studies, processing of the raw data is often needed to generate research-grade exposure and outcome variables, a process called *phenotyping*. This process can be rule-based or involve probabilistic algorithms (e.g., see Kirby et al. [6] and Hubbard et al. [7]). However, quality issues in EHRs (including inaccurate data and data fragmentation [8]) can lead to error-prone phenotyping,

These authors contribute equally to this paper.

resulting in misclassification or measurement error in the derived variables. The use of misclassified variables poses a major threat to the reproducibility [9] of EHR-based discoveries and undermines its promise in expanding the horizons of traditional clinical research.

Exposure-dependent phenotyping errors (a form of differential misclassification) are of particular concern in EHR-based studies. Association studies that utilize derived outcomes subject to exposure-dependent misclassification have been shown to suffer from reduced statistical power [10], inflated type I errors [11], and biased association parameter estimates [12, 13]. This can occur when certain patient groups have more information in their health records (due to a larger number of visits at the same institution or better documentation of health status) to enable accurate phenotyping. Phenotyping errors can therefore substantially affect conclusions on the comparative effectiveness of interventions.

Manual chart review remains the gold standard for validating EHR-derived phenotypes and ensuring high-quality data for clinical research [14]. Due to the massive size of EHR databases combined with budget and time constraints, performing manual chart review for the entire sample is typically infeasible, and only a subset can be validated. Several methods have been developed for bias reduction and improved efficiency in the presence of outcome misclassification in semi-supervised settings [15–17], where the true outcome is known for a small, validated subset of the sample. These methods aim to produce unbiased estimates of the parameters of interest using a labeled sample, while the unlabeled data are used to improve precision, sometimes with the aid of surrogate variables. However, the quality of the validated sample is often overlooked. Investigators have the opportunity to consider alternative designs for selecting validation sets, which can improve precision of their association parameter estimates. Sampling designs that select the most informative subjects for chart review (either as a validation set for a semi-supervised method or a standalone sub-sample for estimation) can yield more efficient estimators under a given budget and/or time constraint.

We draw from the literature on algorithmic leveraging to propose sampling designs for validating EHR outcomes that improve statistical efficiency. Algorithmic leveraging is a sampling process that utilizes statistical leverage (or "importance") scores to identify the most influential observations for a given model [18]. It has traditionally been applied to large-scale matrix problems (including least squares approximation [19] and low-rank matrix approximation [20]) and was motivated by the need to reduce computational cost with limited computing resources in the large n and/or large p setting. Rather than analyzing the full sample, computations would instead be performed on the selected subsample, and the estimate would approximate that using the full data set. In recent years, algorithmic leveraging has been proposed to improve statistical efficiency (defined using the mean squared error) in the settings of linear regression [18, 21] and logistic regression [22–24].

For the setting of EHR data with binary outcomes, we are interested in fitting a logistic regression model of the gold-standard outcome, y, conditional on a feature vector $\boldsymbol{x}$. Therefore, y must be uncovered through chart review. Most existing efficient sampling schemes for logistic regression either assume that the gold-standard outcome is fully observed [23], or missing entirely [24], neither of which is appropriate for our setting. Notably, while we do not observe the gold-standard outcome y, we have auxiliary information through the observed, EHR-derived phenotype, s. Alternatively, s may be a proxy or auxiliary variable that is associated with the outcome, but would not be included in the association model, as the resulting coefficients (e.g., the effect sizes of risk factors) would not have their intended interpretations. Pepe et al. [25]. developed sampling weights that depend on auxiliary data, but these are specific to an estimator that solves the mean score equation using both the validation and non-validation set. We seek to focus on the estimator based on the validation set only, which can then be incorporated into other ad-hoc approaches using the non-validation sample.

In this paper, we propose an optimal surrogate-assisted sampling scheme for validating EHR outcomes that makes use of both the covariates, $\boldsymbol{X}$, and the surrogate outcome, $\boldsymbol{s}$. We aim to bridge the gap between current semi-supervised methods (which ignore design considerations in selecting validation sets) and existing subsampling algorithms for logistic regression. The proposed sampling design aims for the subsample-based estimator to approximate the full data maximum-likelihood estimator (MLE) (the estimate if we had observed y for all individuals). This is achieved by applying the A-optimality criterion [26] to the asymptotic approximation error. Our proposed sampling weights are A-optimal for the logistic regression model of $\boldsymbol{y}$ regressed on $\boldsymbol{X}$ given the informative surrogate outcome $\boldsymbol{s}$. We study the strengths of each set of weights under different sample size constraints and accuracy levels of the phenotype and apply them to an EHR data set from Kaiser Permanente Washington (KPWA) to study risk factors for second breast cancer events (SBCE) in women with a personal history of breast cancer.

This article is organized as follows. In Section 2, we describe the general sampling framework for logistic regression and introduce the proposed surrogate-assisted sampling weights. In Section 3, we perform a simulation study to investigate the performance of the weights under high and low sensitivity/specificity of the surrogate phenotype. Finally, in Section 4, we apply the various weighting approaches to the KPWA study of second breast cancers. A brief discussion is offered in Section 5.

2 | Methods

2.1 | Setting and Logistic Regression

To illustrate our methods, we denote the full data matrix for n subjects as $\mathcal{F}_n = (\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{s})$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$ is the covariate matrix, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is the vector of gold-standard outcomes and $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)^T$ is the vector of surrogate outcomes, representing the EHR-derived phenotype or a proxy variable associated with the outcome. We assume that $y_i, s_i \in \{0, 1\}$, and $\boldsymbol{x}_i \in \mathfrak{R}^p$.

Our interest is in estimating the coefficient vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, which belongs to a compact set in $\mathfrak{R}^p$, from

the following logistic regression model of y_i conditional on $\boldsymbol{x}_i$,

$$\text{logit}\{p(y_i = 1|\boldsymbol{x}_i)\} = \boldsymbol{x}_i^T\boldsymbol{\beta} \quad (1)$$

where $\text{logit}(a) = \log(a/(1-a))$. Generally, $\boldsymbol{\beta}$ is estimated by maximizing the log-likelihood function with respect to $\boldsymbol{\beta}$, with the maximum likelihood estimator (MLE), defined as

$$\hat{\boldsymbol{\beta}}_{MLE} = \text{argmax}_{\beta\in\Re^p}\ell(\boldsymbol{\beta}) = \text{argmax}_{\beta\in\Re^p} \times\left(\sum_{i=1}^{n}\left[y_i\log\left\{p_i(\boldsymbol{\beta})\right\} + (1-y_i)\log\left\{1-p_i(\boldsymbol{\beta})\right\}\right]\right) \quad (2)$$

where $p_i(\boldsymbol{\beta}) = p(y_i = 1|\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})/\left\{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}$. A numeric solution to the above problem is usually obtained through iterative methods such as Newton-Raphson or Fisher scoring.

We consider the EHR setting where only $(\boldsymbol{x}_i, s_i)$ is observed for the full sample, and y_i must be uncovered through chart review. As it is time- and cost-prohibitive to validate outcomes for the entire data set, we only uncover y_i for $r \ll n$ individuals. Our goal is to select the most informative sample of r individuals for our model.

2.2 | General Sampling Scheme for Outcome Validation in Logistic Regression

In Algorithm 1, we outline a general subsampling algorithm for logistic regression models with outcome validation. When $r \ll n$ data points are sampled with replacement for validation and analysis, the observations must be weighted by the inverse of their respective sampling probabilities, π_i^*, when fitting the model. We annotate variables with a (*) if they correspond to individuals who have been selected into the validation sample. Furthermore, we denote the solution to the reweighted score equation based on the subsample as $\tilde{\boldsymbol{\beta}}$.

ALGORITHM 1 | General subsampling algorithm for outcome validation.

1. Sample $r(\ll n)$ data points with replacement from the full data set with sampling probabilities $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$. For selected observations, uncover the true outcome, y_i, and denote the sampled data points as $\boldsymbol{O}_i^* = (\boldsymbol{x}_i^*, s_i^*, y_i^*, \pi_i^*)$, for $i = 1, \ldots, r$.
2. Maximize the weighted pseudo log-likelihood to obtain the subsample-based estimate $\tilde{\boldsymbol{\beta}}$.

$$\tilde{\boldsymbol{\beta}} = \underset{\beta\in\Re^p}{\text{argmax}}\ell(\boldsymbol{\beta}) = \underset{\beta\in\Re^p}{\text{argmax}} \left(\sum_{i=1}^{r}\frac{1}{\pi_i^*}\left[y_i^*\log\left\{p_i^*(\boldsymbol{\beta})\right\} + (1-y_i^*)\log\left\{1-p_i^*(\boldsymbol{\beta})\right\}\right]\right),$$

where $p_i^*(\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^{*T}\boldsymbol{\beta})/\left\{1+\exp(\boldsymbol{x}_i^{*T}\boldsymbol{\beta})\right\}$.

This is equivalent to finding the solution to the following reweighted score equation:

$\dot{\ell}^* = \frac{1}{r}\sum_{i=1}^{r}\frac{\left\{y_i^*-p_i^*(\tilde{\boldsymbol{\beta}})\right\}\boldsymbol{x}_i^*}{\pi_i^*} = \boldsymbol{0}$.

A numeric solution can be obtained through iterative methods such as Newton-Raphson, which performs iterations of the following formula until convergence of $\tilde{\boldsymbol{\beta}}$.

$$\tilde{\boldsymbol{\beta}}^{(t+1)} = \tilde{\boldsymbol{\beta}}^{(t)} + \left[\sum_{i=1}^{r}\frac{p_i^*(\tilde{\boldsymbol{\beta}}^{(t)})\left\{1-p_i^*(\tilde{\boldsymbol{\beta}}^{(t)})\right\}\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right]^{-1}\sum_{i=1}^{r}\frac{\left\{y_i^*-p_i^*(\tilde{\boldsymbol{\beta}}^{(t)})\right\}\boldsymbol{x}_i^*}{\pi_i^*}$$

In the setting where the gold-standard outcome is observed for all individuals, Wang et al. [23] proved that $\tilde{\boldsymbol{\beta}}$ is consistent for $\hat{\boldsymbol{\beta}}_{MLE}$, conditional on the full data matrix $\mathcal{F}_n$, as $n \to \infty$ and $r \to \infty$ (see section S.1.1 of Wang et al. [23]), under certain regularity conditions on the covariate distribution. Furthermore, they showed that as $n \to \infty$ and $r \to \infty$ (where n increases at a faster rate, such that $r = o(n)$, or equivalently $r/n \to 0$) and conditional on $\mathcal{F}_n$,

$$r^{1/2}\boldsymbol{V}^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I})$$

where $\boldsymbol{V} = \boldsymbol{M}_X^{-1}\text{Var}(\boldsymbol{\psi}_i|\mathcal{F}_n)\boldsymbol{M}_X^{-1} = O_p(r^{-1})$, $\boldsymbol{\psi}_i = \left\{y_i^* - p_i^*(\tilde{\boldsymbol{\beta}})\right\}\boldsymbol{x}_i^*/(n\pi_i^*)$ and $\boldsymbol{M}_x = \sum_{i=1}^n p_i(\hat{\boldsymbol{\beta}}_{MLE})\left\{1-p_i(\hat{\boldsymbol{\beta}}_{MLE})\right\}\boldsymbol{x}_i\boldsymbol{x}_i^T/n$.

While this result was derived for the setting where the gold-standard outcome y_i is observed, this also holds true for the weighted estimator $\tilde{\boldsymbol{\beta}}$ in Algorithm 1, as the final weighted analysis is performed on observations, where y_i has been validated, and Wang et al.'s (2018) result is agnostic to the specific form of π_i (apart from requiring that $\sum_{i=1}^n\pi_i = 1$).

2.3 | Prior Work on Optimal Sampling Weights

A simple choice for the sampling weights in Algorithm 1 is $\pi_i = 1/n$. This is known as *uniform sampling* (UNI). Another choice is uniform sampling stratified by the outcome, that is, $\pi_i = 1/(2n_i)$ with n_0 and n_1 being the number of subjects with outcome $y = 0$ and $y = 1$, respectively. This is known as *case-control sampling* and is commonly used when the outcome is unbalanced. However, these simple sampling weights may not be the "optimal", choice in the sense that $\tilde{\boldsymbol{\beta}}$ may be estimated with greater precision under alternative subject-specific weights. Motivated by large-scale data problems, Wang et al. [23] derived optimal subsampling procedure motivated from the A-optimality criterion (OSMAC). We denote these weights as $\pi_{i,OSMAC(y)}$ (reflecting that they are optimal when y_i is observed for the full sample),

$$\pi_{i,OSMAC(y)} = \frac{|y_i - p_i(\hat{\boldsymbol{\beta}}_{MLE})|||\boldsymbol{M}_x^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^n|y_j - p_j(\hat{\boldsymbol{\beta}}_{MLE})|||\boldsymbol{M}_x^{-1}\boldsymbol{x}_j||} \quad (3)$$

where $||\boldsymbol{v}||$ is the euclidean norm of a vector $\boldsymbol{v}$ (i.e., $||\boldsymbol{v}|| = (\boldsymbol{v}^T\boldsymbol{v})^{1/2}$).

Wang et al. [23] proposed a two-step approach, in which a small sample is first randomly selected to obtain a pilot estimate of $\hat{\boldsymbol{\beta}}_{MLE}$. This enables us to characterize the relationship between y_i and $\boldsymbol{x}_i$ to determine which observations are "surprising" given their expected value. $\pi_{i,OSMAC(y)}$ are then calculated and a more informative subsample is selected for estimating $\tilde{\boldsymbol{\beta}}$.

For the setting where y_i is unknown and only $\boldsymbol{x}_i$ is observed, Zhang et al. [24] proposed an optimal sampling under measurement constraints (OSUMC), which does not require known outcome values. We denote these weights as $\pi_{i,OSUMC}$, and

$$\pi_{i,OSUMC} = \frac{\sqrt{p_i(\hat{\boldsymbol{\beta}}_{MLE})\left\{1-p_i(\hat{\boldsymbol{\beta}}_{MLE})\right\}}||\boldsymbol{M}_x^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^n\sqrt{p_j(\hat{\boldsymbol{\beta}}_{MLE})\left\{1-p_j(\hat{\boldsymbol{\beta}}_{MLE})\right\}}||\boldsymbol{M}_x^{-1}\boldsymbol{x}_j||} \quad (4)$$

where again $\hat{\boldsymbol{\beta}}_{MLE}$ is estimated from a pilot sample. Note that unlike the weights given by Wang et al. [23] ($\pi_{i,\text{OSMAC(y)}}$), which seek to identify individuals for whom the observed outcome y_i is unexpected or extreme given their $\boldsymbol{x}_i$, the weights proposed by Zhang et al. [24] prioritize individuals with fitted probabilities closest to 0.5 (which is equidistant between the two possible values of y_i, $\{0, 1\}$). Intuitively, this means that $\pi_{i,\text{OSUMC}}$ gives greatest weight to individuals for whom the model is most uncertain about their outcome. This approach, known as uncertainty-based sampling, is rationally sound in the absence of any additional information to guide sample selection. However, it relies critically on the assumption that the fitted probability is a good fit to the data. In the next section, we propose weights that use the surrogate outcome s to offer additional insight on which individuals will have a y_i value that is "surprising" given their fitted probabilities $p_i(\hat{\boldsymbol{\beta}}_{MLE})$.

2.4 | Surrogate-Assisted Sampling for Outcome Validation

Having reviewed the "optimal" weights for the setting where the true outcome y_i is observed for everyone, and alternatively, the setting where y_i is missing (with only $\boldsymbol{x}_i$ being available), we now turn our attention to the setting where y_i is potentially misclassified, and a surrogate outcome s_i is observed along with $\boldsymbol{x}_i$ for all individuals. This is often encountered in work with EHR data, where phenotyping errors may occur in the use of automated algorithms. An immediate approach is to simply use the surrogate instead of the true outcome in the OSMAC weights, denoted as OSMAC(s). However, the performance of this subsampling strategy depends on the accuracy of the surrogate outcome, which is difficult to evaluate before a validation sample is drawn. In this section, we propose an Optimal Subsampling with Surrogate-Assisted Two-step procedure (OSSAT) to be used in place of the targeted optimal weights when the true outcome $\boldsymbol{y}$ were observed (e.g., OSMAC(y)). This novel sampling approach makes the best use of the surrogate outcome and is demonstrated to perform better than OSMAC(s) when the accuracy of the surrogate outcome is not satisfactory. By making use of both s_i and $\boldsymbol{x}_i$ to identify influential observations for estimation of $\tilde{\boldsymbol{\beta}}$, we seek to achieve greater efficiency compared to the use of $\boldsymbol{x}_i$ alone (as is done with $\pi_{i,\text{OSUMC}}$).

2.4.1 | Surrogate-Substituted Weights

The surrogate-substituted OSMAC weight simply uses the surrogate outcome $\boldsymbol{s}$ instead of the true outcome $\boldsymbol{y}$ in $\pi_{i,\text{OSMAC(y)}}$, namely

$$\pi_{i,OSMAC(s)} = \frac{|s_i - p_i(\hat{\boldsymbol{\beta}}_{MLE})|||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^{n}|s_j - p_j(\hat{\boldsymbol{\beta}}_{MLE})|||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_j||} \tag{5}$$

where $\hat{\boldsymbol{\beta}}_{MLE}$ is obtained using the full data with s being the response. The weight $\pi_{i,OSMAC(s)}$ is expected to approximate $\pi_{i,OSMAC(y)}$ well when s_i agrees with y_i, but may be much different from $\pi_{i,OSMAC(y)}$ when they disagree. The performance of OSMAC(s) is hence highly dependent on the accuracy of the surrogate outcome. The misclassification of the surrogate versus the true outcome is characterized by sensitivity and specificity, that is, $se = Pr(s = 1|y = 1)$ and $sp = Pr(s = 0|y = 0)$. In the following derivations, we consider the non-differential misclassification setting, where $s \perp \boldsymbol{x}|y$. This implies that se, sp are common for each subject and do not depend on covariates, that is, $se = se_i = Pr(s_i = 1|y_i = 1) = Pr(s_i = 1|y_i = 1, \boldsymbol{x}_i)$ and $sp = sp_i = Pr(s_i = 0|y_i = 0) = Pr(s_i = 0|y_i = 0, \boldsymbol{x}_i), i = 1, \cdots, n$. In the numerical analysis, we will demonstrate that this assumption is robust to differential misclassification, where se, sp depend on covariates $\boldsymbol{x}$.

2.4.2 | Optimal Surrogate-Augmented Weights

In this section, we propose a new method to compute sampling weights by incorporating surrogate information. Similar to Wang et al. [23], we consider A-optimality for asymptotic variance of $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. It turns out that the asymptotic variance of $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ given the feature matrix $\boldsymbol{X}$ can be decomposed as

$$\begin{aligned}\text{Var}\{(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})|\boldsymbol{X}\} = \text{Var}\{(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})|\boldsymbol{X}\} + \text{Var}\{(\hat{\boldsymbol{\beta}}_{MLE} - \boldsymbol{\beta})|\boldsymbol{X}\} \\ + 2\text{E}\{(\hat{\boldsymbol{\beta}}_{MLE} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})|\boldsymbol{X}\}\end{aligned} \tag{6}$$

where $\boldsymbol{\beta}$ is the true parameter and $\hat{\boldsymbol{\beta}}_{MLE}$ is obtained from on the full data. The second term is simply the asymptotic variance of $\hat{\boldsymbol{\beta}}_{MLE}$, which is not related to the sampling weights π_i. Moreover, it can be shown that the third term does not depend on π_i (see details in Sections S1.2). Consequently, minimizing the asymptotic variance of $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ with respect to π_i is equivalent to minimizing the asymptotic variance of $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})$.

If y_i is known, one can proceed to derive the A-optimal sampling weights, $\pi_{i,\text{OSMAC(y)}}$, that minimize the asymptotic variance of $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})$ by minimizing its trace, as is done in Wang et al. [23]. However, if y_i is not known, we propose applying the law of total variance, also known as the variance decomposition formula. This partitions the sample space for $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})$ over the distribution of one or more components. Zhang et al. [24] first applied this formula for a single component, $\boldsymbol{y}$. We propose to additionally apply it for $\boldsymbol{s}$ as follows,

$$\begin{aligned}\text{Var}\{(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})|\boldsymbol{X}\} &= \text{E}\{\text{Var}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{X})|\boldsymbol{s}, \boldsymbol{X}\} \\ &+ \text{E}\{\text{Var}(\text{E}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{X}]|\boldsymbol{s}, \boldsymbol{X})|\boldsymbol{X}\} \\ &+ \text{Var}\{\text{E}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s}, \boldsymbol{X})|\boldsymbol{X}\}\end{aligned} \tag{7}$$

Proceeding with minimization of trace$[\text{Var}\{(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})|\boldsymbol{X}\}]$, and assume $p_i = Pr(y_i|\boldsymbol{x}_i), p_i^s = Pr(y_i|s_i, \boldsymbol{x}_i)$, we find the following.

Proposition 1. *If the subsampling probability in Algorithm 1 is set to*

$$\pi_{i,\text{OSSAT}} = \frac{\sqrt{\{p_i^s - 2p_i^s p_i + p_i^2\}}||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^{n}\sqrt{\{p_j^s - 2p_j^s p_j + p_j^2\}}||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_j||} \tag{8}$$

then the asymptotic approximation error of $\tilde{\boldsymbol{\beta}}$, *defined as* trace$[\text{Var}\{(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})|\boldsymbol{X}\}]$, *will attain its minimum* (*see Section S1.3 for proof*).

As the true outcome y is required in the evaluation of the subsampling probability $\pi_{i,OSSAT}$ in real applications, a two-step algorithm is designed to get the subsample for validation as outlined in Algorithm 2.

ALGORITHM 2 | Surrogate-augmented subsampling for outcome validation.

1. **Step 1 (Pilot)**: Sample r_1 observations by Algorithm 1, using the surrogate-based balanced stratified sampling weight $\pi_{i,CC(s)} = 1/(2n_{si})$, in which n_{s0} and n_{s1} are the number of subjects with surrogate $s = 0$ and $s = 1$, respectively. For sampled observations, chart review is performed to uncover y. Pilot estimates for $\hat{p}_i^s$ and $\hat{\boldsymbol{\beta}}_{MLE}$ are calculated using the r_1 observations.

2. **Step 2 (Optimal sampling)**: Pilot estimates from Step 1 are plugged into the surrogate-augmented optimal weights $\pi_{i,\text{OSSAT}}$ by Proposition 1. These weights are used to sample r_2 individuals with replacement from the original n data points, and we collect the chart-reviewed outcome for selected subjects. The final estimate $\tilde{\boldsymbol{\beta}}$ proceeds through weighted estimation combining samples from Steps 1 and 2:

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\Re^p}{\text{argmax}} \frac{1}{r_1 + r_2}\left(\sum_{i\in step1\ set} \frac{1}{\pi^*_{i,CC(s)}} \ell_i^*(\boldsymbol{\beta}) + \sum_{i\in step2\ set} \frac{1}{\pi^*_{i,OSSAT}} \ell_i^*(\boldsymbol{\beta}) \right),$$

where $\ell_i^*(\boldsymbol{\beta}) = y_i^* \log\left\{p_i^*(\boldsymbol{\beta})\right\} + (1 - y_i^*)\log\left\{1 - p_i^*(\boldsymbol{\beta})\right\}$, and $p_i^*(\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^{*T}\boldsymbol{\beta})/\left\{1 + \exp(\boldsymbol{x}_i^{*T}\boldsymbol{\beta})\right\}$.

We remark that the surrogate-based balanced stratified sampling (CC(s)) used in the pilot step is constructed by utilizing $\boldsymbol{s}$ in place of $\boldsymbol{y}$ in the case-control sampling (CC) as $\boldsymbol{y}$ is unknown.

For Algorithm 2, we need to estimate the probability p_i^s. With a pilot sample, a simple approach is to fit a logistic regression model of y_i on $\{s_i, \boldsymbol{x}_i\}$, yet this is a misspecified model given that $y|\boldsymbol{x}$ follows a logistic regression model. A better approach is via the Bayesian rule and pilot estimation of sensitivity and specificity. Specifically, under non-differential misspecification when sensitivity and specificity are not dependent on $\boldsymbol{x}$, we have the following reformulation

$$Pr(y = 1|s, \boldsymbol{x}) = \frac{Pr(s|y = 1)Pr(y = 1|\boldsymbol{x})}{Pr(s|y = 1)Pr(y = 1|\boldsymbol{x}) + Pr(s|y = 0)Pr(y = 0|\boldsymbol{x})},$$

hence

$$Pr(y_i = 1|s_i = 1, \boldsymbol{x}_i) = \frac{se \times p_i}{se \times p_i + (1 - sp)(1 - p_i)}$$

$$Pr(y_i = 1|s_i = 0, \boldsymbol{x}_i) = \frac{(1 - se) \times p_i}{(1 - se) \times p_i + sp \times (1 - p_i)}$$

Working with the 2×2 contingency table, we have

$$se = \frac{(1 - p) - p \times npv/(1 - npv)}{p((1 - ppv)/ppv - npv/(1 - npv))},$$
$$sp = \frac{p - (1 - p) \times ppv/(1 - ppv)}{(1 - p)((1 - npv)/npv - ppv/(1 - ppv))} \quad (9)$$

where ppv is the positive predicted value (PPV) and npv is the negative predicted value (NPV), both estimated from the pilot sample, that is $ppv = Pr(y = 1|s = 1)$ and $npv = Pr(y = 0|s = 0)$, and $p = Pr(s = 1)$ is the prevalence of y estimated from s using the full data.

Remark 1. Convergence issues and biased model estimates may result when fitting the models for $p(y_i|\boldsymbol{x}_i)$ in pilot step 1 samples with rare events, preventing successful construction of the weights. To overcome this, we recommend using the Firth adjustment to the weighted pseudo log-likelihood in small samples [27].

Remark 2. Calculating the sensitivity and specificity from the observed ppv, npv and p may not be successful as the quantities in Equation (9) may fall out of the meaningful range of 0.5–1 in cases such as rare events and too small pilot sample. In these extreme cases, we suggest truncating the calculated sensitivity and specificity values within 0.5–1 for numerical stability.

Remark 3. Intuitively, if p_i^s can be estimated with precision in the first step of sampling, then we can expect $\pi_{i,\text{OSSAT}}$ to perform similarly to $\pi_{i,\text{OSMAC(y)}}$. This is because for each subject, $\sqrt{\left\{p_i^s - 2p_i^s p_i + p_i^2\right\}}$, will approximate $\sqrt{E\left\{(y_i - p_i)^2|s_i, \boldsymbol{x}_i\right\}}$ as $r \to \infty$, which by Jensen's inequality is $\geq E\left\{\sqrt{(y_i - p_i)^2}|s_i, \boldsymbol{x}_i\right\} = E\left\{|y_i - p_i||s_i, \boldsymbol{x}_i\right\}$. Since this relationship holds point-wisely for each $s_i, \boldsymbol{x}_i$, we argue that some of the ordering of observations according to their informativeness is preserved, not accounting for random error.

3 | Simulation Study

We evaluate how well each set of surrogate-augmented weights approximate the optimal weights $\pi_{i,\text{OSMAC(y)}}$, as well as their ability to improve the precision of $\tilde{\boldsymbol{\beta}}$, the subsample-based estimate of $\boldsymbol{\beta}$ in the logistic regression model of y_i on $\boldsymbol{x}_i$. We consider and compare model estimates using the optimal weights $\pi_{i,\text{OSMAC(y)}}$ which requires the true outcome, our proposed surrogate-augmented optimal weights $\pi_{i,\text{OSSAT}}$, which only requires surrogate outcome, and other alternative approaches including $\pi_{i,\text{CC(s)}}$, $\pi_{i,\text{OSUMC}}$, $\pi_{i,\text{OSMAC(s)}}$, and $\pi_{i,\text{UNI}}$.

3.1 | Simulation Setting

We consider multiple covariate distributions as described in [23] and [24], with the number of covariates as 7 and the distributions including:

- **zeroMean**. $\boldsymbol{x}$ is multivariate normal with constant variance, defined as $MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5^{i\neq j}$, such that the diagonals of Σ are equal to 1, and the off-diagonals are equal to 0.5.
- **unequalVar**. $\boldsymbol{x}$ is multivariate normal with unequal variances, that is $\boldsymbol{x} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}_{ii} = 1/i^2$ for $i = 1, \ldots, 7$, and the off-diagonal entries equal to $\Sigma^*_{ij} = 0.5$ for $i \neq j$.

- **rareEvent**. $\boldsymbol{x}$ is multivariate normal, centered away from 0, which induces extreme imbalance in the outcomes ($\sim 5\%$ cases). $\boldsymbol{x} \sim MVN(-\boldsymbol{1.6}, \boldsymbol{\Sigma})$, $\Sigma_{ij} = 0.5^{i \neq j}$.
- **mixNormal**. $\boldsymbol{x}$ follows a bimodal distribution that is the mixture of two multivariate normal distributions $0.5N(\boldsymbol{1}, \boldsymbol{\Sigma})$ and $0.5N(-\boldsymbol{1}, \boldsymbol{\Sigma})$, $\Sigma_{ij} = 0.5^{i \neq j}$.
- $\boldsymbol{T}_3$. $\boldsymbol{x}$ follows a multivariate t distribution with degrees of freedom 3, $t_3(\boldsymbol{0}, \boldsymbol{\Sigma})/10$. For this case, the distribution of $\boldsymbol{x}$ has a heavy tail.
- **Exp**. Each component of $\boldsymbol{x}$ follows an exponential distribution with a rate parameter of 2. The covariates are uncorrelated in this setting.

We then generate the true outcome y_i by $p_i = expit(\boldsymbol{x}_i^T \boldsymbol{\beta})$, and $\boldsymbol{\beta}$ has true values equal to $(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$ with the first element being the intercept. For the scenario that $\boldsymbol{x} \sim$ **Exp**, we set the intercept as -0.5 and the prevalence of y is close to 0.75. We assume differential misclassification for the surrogate outcome, that is, the sensitivity and specificity depend on the covariates. Specifically, we consider a scenario with less misclassification, that is, $(se, sp) = (0.1I(x_1 < c_1) + 0.8, 0.04I(x_1 < c_1) + 0.95)$, and a scenario with more misclassification, that is, $(se, sp) = (0.1I(x_1 < c_1) + 0.6, 0.05I(x_1 < c_1) + 0.9)$, where c_1 is the 30% quantile of x_1.

To study the performance of the proposed sampling designs on model efficiency, we perform $S = 500$ replications under each covariate distribution. For each replicate, a data set of $n = 10,000$ is generated, and the total subsample size to be validated is $r = \{800, 1000, 1200, 1400, 1600\}$. For the two-step approaches only, r includes a pilot step 1 sample of size $r_1 = \{200, 600\}$ sampled by surrogate-based balanced stratified sampling, and the remainder is selected in step 2 with the more informative weights (i.e., $r = r_1 + r_2$). For each sample size and covariate distribution setting, we calculate the empirical mean-squared error (as compared to the true model parameters) as follows,

$$\text{MSE}_{\beta} = \frac{1}{S}\sum_{s=1}^{S} ||\tilde{\boldsymbol{\beta}}_s - \boldsymbol{\beta}||^2$$

where $\boldsymbol{\beta}$ is the true parameter vector, and $\tilde{\boldsymbol{\beta}}_s$ is the estimate from the s^{th} replicate, and S is the number of replications. We also assess the concordance of the proposed OSSAT weights, the surrogate-substitute OSMAC(s) weights versus the OSMAC(y) weights for the various covariate distributions. The correlation between subsampling weights is calculated and compared.

3.2 | Simulation Results

Figure 1 shows the individual estimated weights against their corresponding $\pi_{i,\text{OSMAC(y)}}$ (in log10 scale) with low sensitivity and specificity and $r_1 = 600$. Under perfect concordance between the weights, we would expect the points to fall along the diagonal line. Note that the weights farthest from the origin are of primary interest, as these observations will have the highest probabilities of being included in the second step sample for estimation of the final model. Results show that $\pi_{i,\text{OSSAT}}$ offers closer approximation to $\pi_{i,\text{OSMAC(y)}}$ on average, as points are scattered more evenly about the line $y = x$. However, with a small step 1 sample, the $\pi_{i,\text{OSSAT}}$ weights are estimated with lower precision, and so have greater variability compared to $\pi_{i,\text{OSMAC(s)}}$ (which do not require a step 1 sample), particularly if the event is rare. We see in the lower quadrants of each panel from Figure 1 that if sensitivity and specificity are low, $\pi_{i,\text{OSSAT}}$ is preferred in both small and large samples, since they are more likely to lead to selection of similarly informative observations as if we had used $\pi_{i,\text{OSMAC(y)}}$, as there is stronger concordance among the larger weights. On the other hand, the substitution weights, $\pi_{i,\text{OSMAC(s)}}$ follow a near random distribution when sensitivity and specificity are low, particularly if the event is rare.

Figures 2 and 3 show the mean squared error (MSE) results for $\tilde{\boldsymbol{\beta}}$ using various weights, including $\pi_{i,\text{OSMAC(y)}}$, $\pi_{i,\text{OSSAT}}$, $\pi_{i,s\text{OSMAC(s)}}$, $\pi_{i,\text{OSUMC}}$, $\pi_{i,\text{CC(s)}}$, and $\pi_{i,\text{UNI}}$. Results are displayed for $r_1 = 200$, with results for $r_1 = 600$ included in Figures S1 and S2 in the Supporting Information.

As expected, the weighting method by Wang et al. [23], OSMAC(y), achieves optimal efficiency in all settings. Surrogate-based balanced stratified sampling (CC(s)) and uniform sampling (UNI) are the least efficient in most setting, while UNI provides similar performance to the informative sampling methods OSMAC(s), OSUMC, and OSSAT when the surrogate outcome is less informative to the true outcome under the Exp data set. The proposed weights $\pi_{i,\text{OSSAT}}$ achieve the smallest MSEs compared to all other weights. The advantage of OSSAT over OSMAC(s) is more evident when sensitivity and specificity are low (right panels), as OSSAT combines the true outcome from the pilot sample and the surrogate outcome from the full data, while OSMAC(s) relies on good surrogate outcome. Also, as OSUMC does not depend on the surrogate outcome, its disadvantage over OSMAC(s) becomes smaller or even reverses when sensitivity and specificity become low. These observations hold true with $r_1 = 600$ as shown in the Supplementary Figures.

Furthermore, we observe that OSSAT always results in a similar or slightly lower MSE than OSUMC. This is expected since OSUMC and OSSAT are both derived using the law of total variance, and under zero correlation between s_i and y_i, $p(y_i|s_i, \boldsymbol{x}_i)$ will reduce to $p(y_i|\boldsymbol{x}_i)$ and $\pi_{i,\text{OSSAT}}$ will reduce to $\pi_{i,\text{OSUMC}}$. Thus, under low sensitivity and specificity, the surrogate augmented weights will offer similar or slightly more information compared to weights that utilize $\boldsymbol{x}_i$ only, while the surrogate substitution weights may underperform relative to the weights that use $\boldsymbol{x}_i$ only. Additionally, for the six sampling methods discussed in the manuscript, the finite sample bias is small, and there are little differences between the methods. The finite sample variance makes up the majority of MSE, and it follows the same pattern as the MSE. Additional simulation results are shown in Sections S2 of the Supplement.

4 | Application to BRAVA Study

Here, we apply the candidate weights to an EHR data set from the BRAVA study conducted at Kaiser Permanente Washington (KPWA) [28], which studied risk factors for second breast cancer events (SBCE) in women with a personal history of breast cancer. The data consists of 3152 women diagnosed with primary stage

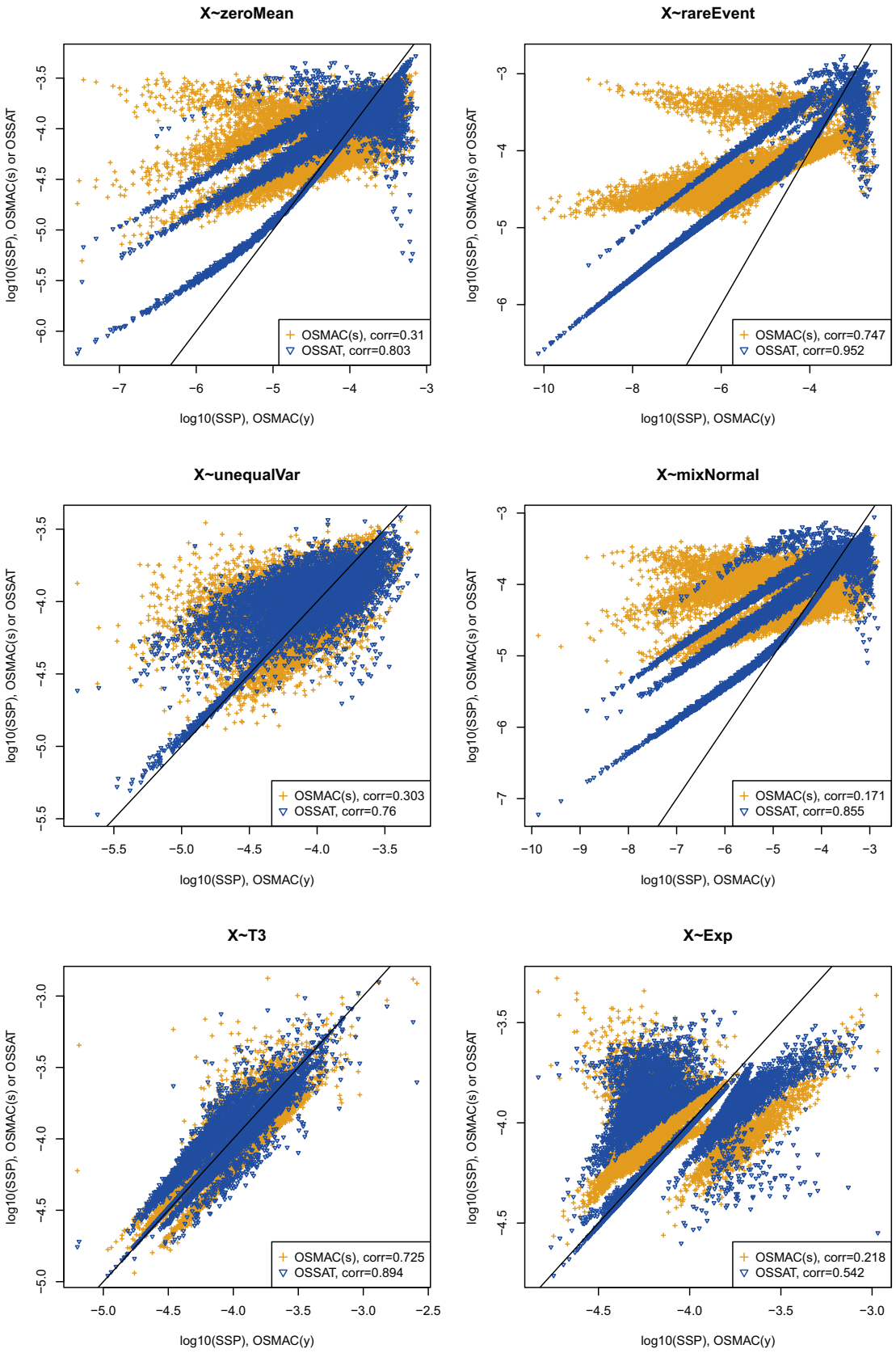

FIGURE 1 | Concordance plots of $\pi_{i,\text{OSSAT}}$ and $\pi_{i,\text{OSMAC(s)}}$, compared to $\pi_{i,\text{OSMAC(y)}}$, under the 6 covariate distributions with $r_1 = 600$ (for $\pi_{i,\text{OSSAT}}$) and low sensitivity and specificity $(se, sp) = (0.1I(x_1 < c_1) + 0.6, 0.05I(x_1 < c_1) + 0.9)$, $n = 10,000$. The correlation coefficients are calculated by Spearman's rank correlation.

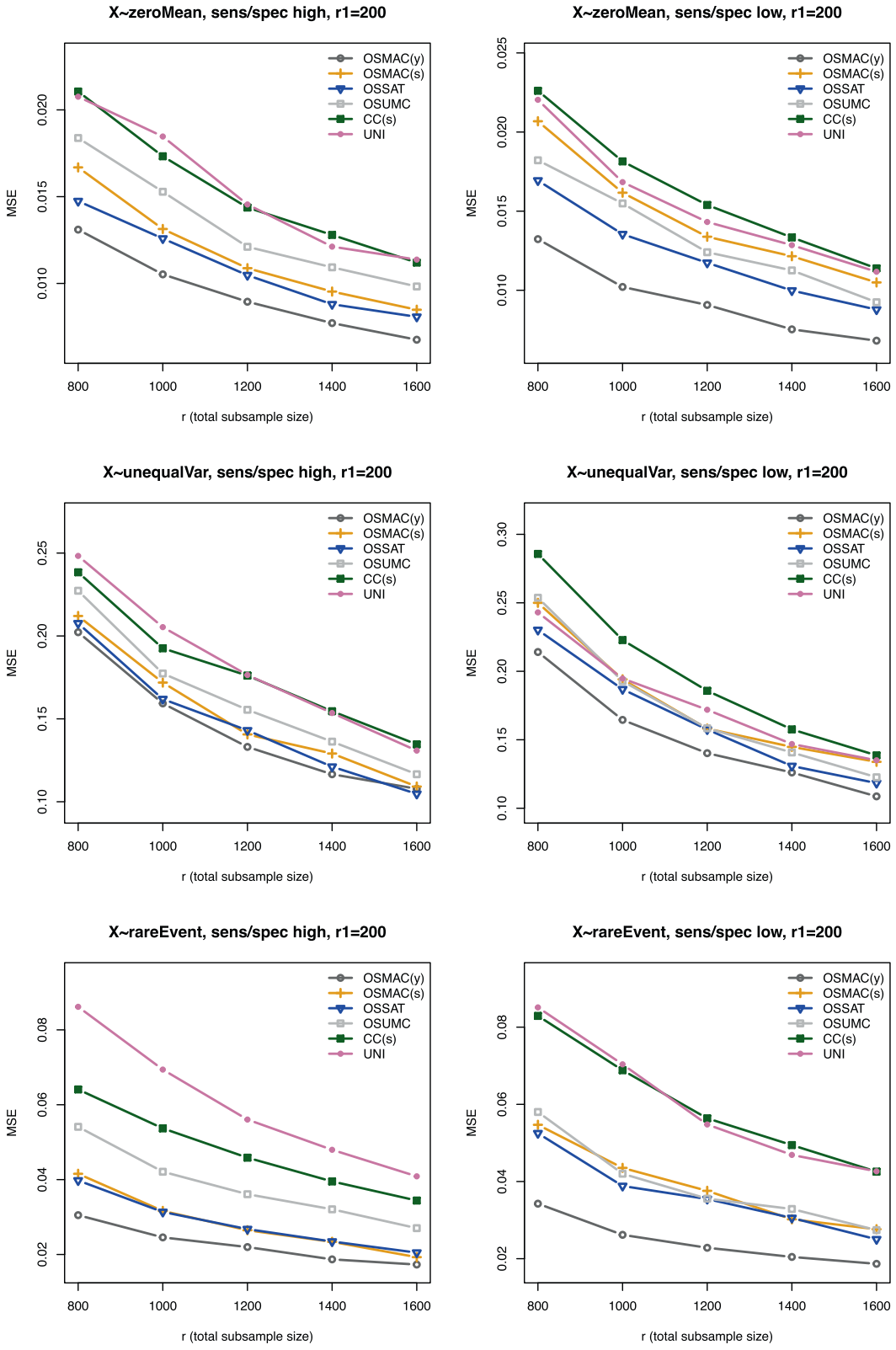

FIGURE 2 | Empirical mean squared error of $\tilde{\beta}$ using different sampling weights, over 500 replicates. Left: high sensitivity and specificity $(se, sp) = (0.1I(x_1 < c_1) + 0.8, 0.04I(x_1 < c_1) + 0.95)$, right: low sensitivity and specificity $(se, sp) = (0.1I(x_1 < c_1) + 0.6, 0.05I(x_1 < c_1) + 0.9)$. The pilot sample size is $r_1 = 200$ with total $n = 10,000$. Three covariate distributions are shown: zeroMean, unequalVar, and rareEvent (5%).

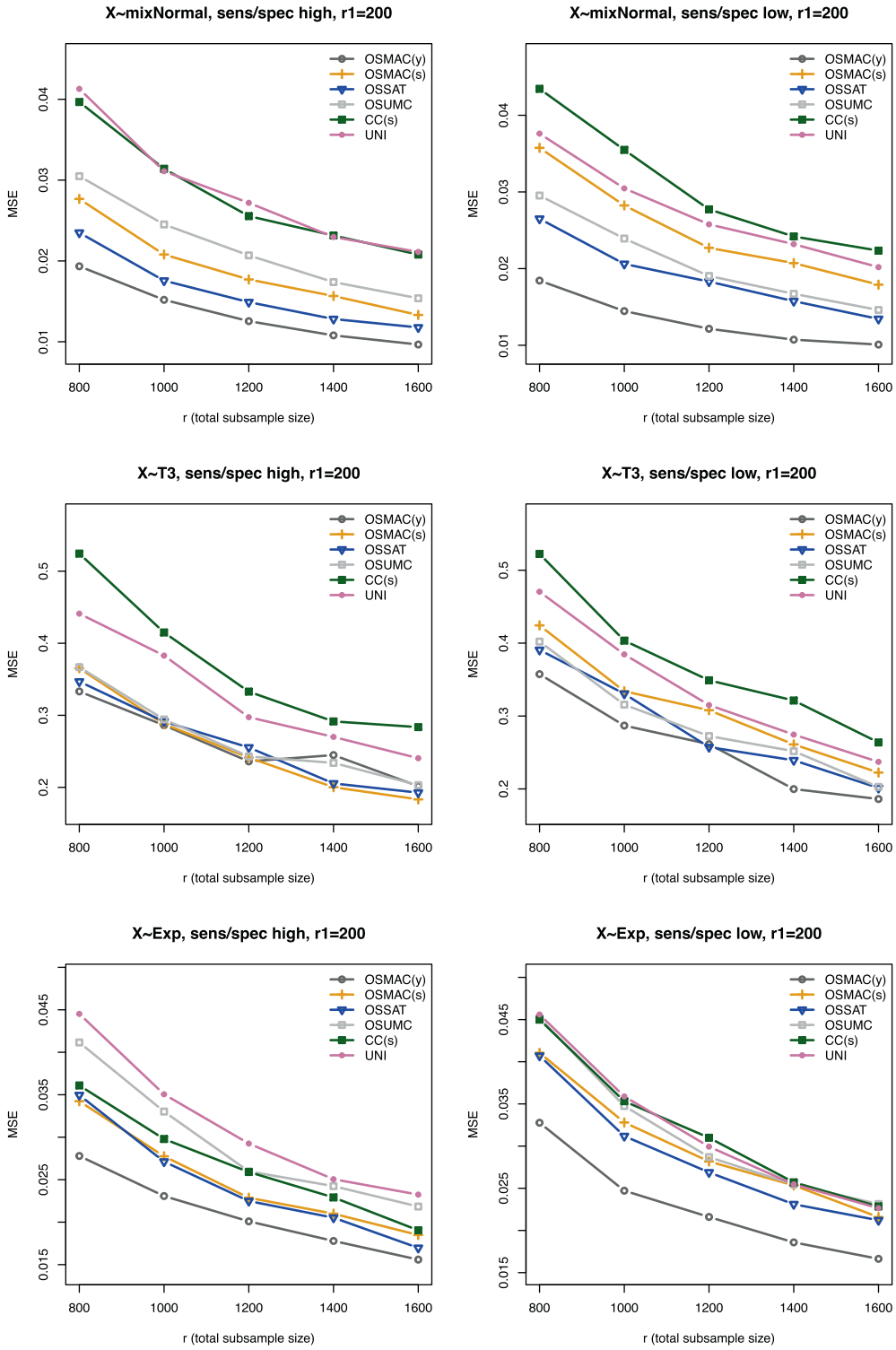

FIGURE 3 | Empirical mean squared error of $\tilde{\beta}$ using different sampling weights, over 500 replicates. Left: high sensitivity and specificity $(se, sp) = (0.1I(x_1 < c_1) + 0.8, 0.04I(x_1 < c_1) + 0.95)$, right: low sensitivity and specificity $(se, sp) = (0.1I(x_1 < c_1) + 0.6, 0.05I(x_1 < c_1) + 0.9)$. The pilot sample size is $r_1 = 200$ with total $n = 10,000$. Three covariate distributions are shown: mixNormal, T_3, and Exp.

I-IIIB invasive breast cancer between 1993 and 2006. This data set is especially useful as we have access to the gold-standard, chart-reviewed outcome of SBCE for all women in the sample, along with a highly specific phenotype developed using classification and regression trees applied to structured EHR and cancer registry data developed by Chubak et al. [29].

We are interested in studying how the risk of developing a SBCE within 2 years of the diagnosis of primary cancer is affected by three risk factors, age at primary breast cancer diagnosis, the stage of the primary cancer in the surveillance, epidemiology, and end results (SEER), and the receptor status of the primary breast cancer (ER/PR-negative vs. ER-positive). Younger age and

later stage at the primary diagnosis are known to associate with elevated risk of SBCE [30, 31]. Hormone therapy can be used to block ER-positive tumors, slowing tumor growth and reducing the risk of recurrent cancer. Patients with ER/PR-negative tumors cannot benefit from this treatment. To illustrate the use of our methods under the logistic regression setting, we include only women with 2 years of follow-up under the assumption of non-informative missingness, such that the total sample size is $n = 2928$. The median age of patients at primary diagnosis is 63 with an interquartile range of $52 - 73$. About 79.2% of the patients are at local stage at the primary diagnosis, while the remainder are at regional stage. The proportion of patients with ER/PR-negative tumors is 13.9%. A total of 375 (12.8%) gold-standard SBCE events are identified. Overall, there are $p = 4$ regression coefficients in the logistic regression model, including the intercept term. The full data MLE is estimated as $\hat{\boldsymbol{\beta}}_{MLE} = [-1.113, -0.009, -0.449, 0.607]$ for intercept, age (year), local vs. regional stage, and negative vs positive ER/PR status, respectively.

We vary the total subsample size from $r = \{250, 350, 450\}$, and the pilot step 1 sample size (for OSSAT and OSUMC) across $r_1 = 80$ and 150. For each sample size setting, we conduct the sampling process independently for $S = 500$ times and calculate the empirical MSE of $\tilde{\boldsymbol{\beta}}$ compared to the full data MLE, $\hat{\boldsymbol{\beta}}_{MLE}$. OSMAC using the true outcome (OSMAC(y)) is used as the oracle benchmark.

The phenotypical algorithm obtains 372 (12.7%) surrogate events, with high specificity (98%) and sensitivity (89%). To better demonstrate the proposed subsampling method when the surrogate sensitivity and specificity are relatively low, we generate a synthetic surrogate based on the true SBCE outcome with differential misclassification. The synthetic surrogate has 85% sensitivity and 85% specificity when age is no more than 54 and has 70% sensitivity and 90% specificity otherwise.

The results of our analysis are shown in Table 1. Uniform sampling with imbalanced outcome data results in a significantly higher empirical MSE compared to all informative sampling approaches. Among these, the surrogate-assisted methods, (OSMAC(s), OSSAT, and CC(s)), achieve lower MSEs by incorporating additional surrogate information. In contrast, OSUMC, which relies solely on covariate information, yields less efficient estimates. Moreover, the performance gap between OSMAC(y) and the surrogate-assisted methods is larger when using the less accurate synthetic surrogate compared to using the true surrogate. This highlights that the efficiency gain through integration of surrogate information into the sampling process depends on the quality of the surrogate.

Both OSSAT and OSMAC(s) use a two-step sampling procedure and use the surrogate information to derive the optimal weights. When the size of the total validation sample is $r = 250$ or 350, OSSAT consistently produces the smallest MSE. The improvement of OSSAT over OSMAC(s) is more pronounced when the synthetic surrogate with lower sensitivity and specificity is used. However, OSMAC(s) provides slightly smaller MSE under two experiments with $r = 450$ validation sample size. Note that, with a fixed pilot sample size r_1, the differences in MSE among informative sampling methods diminish as the total validation sample size r increases to 450. This is likely because, with the increasing

TABLE 1 | Empirical results for estimating the log odds ratios ($\tilde{\boldsymbol{\beta}}$) of risk factors for second breast cancer event using the BRAVA dataset ($n = 2928$). Reported are the empirical MSE of $\tilde{\boldsymbol{\beta}}$ compared to $\hat{\boldsymbol{\beta}}_{MLE}$ under different settings of the surrogate, the pilot step 1 subsample size (r_1) and the total subsample size (r) over 500 replicates. The best result (except OSMAC(y)) under each setting is marked in bold. OSMAC(y): optimal subsampling procedure motivated from the A-optimality criterion using the true outcome; OSMAC(s): OSMAC using the surrogate outcome; OSSAT: optimal subsampling with surrogate-assisted two-step procedure; CC(s): surrogate-based balanced stratified sampling; OSUMC: optimal sampling under measurement constraints; UNI: uniform sampling.

		$r_1 = 80$			$r_1 = 150$		
		r			r		
Setting	**Weights**	**250**	**350**	**450**	**250**	**350**	**450**
True surrogate (high sensitivity and specificity)	OSMAC(y)	0.769	0.525	0.383	0.856	0.528	0.399
	OSMAC(s)	0.919	0.604	0.446	0.936	0.631	**0.428**
	OSSAT	**0.877**	**0.597**	**0.442**	**0.896**	**0.592**	0.456
	CC(s)	1.063	0.608	0.489	0.942	0.673	0.488
	OSUMC	1.230	0.767	0.583	1.069	0.765	0.608
	UNI	1.512	1.061	0.801	1.614	1.152	0.898
Synthetic surrogate (low sensitivity and specificity)	OSMAC(y)	0.784	0.516	0.383	0.877	0.563	0.410
	OSMAC(s)	1.052	0.770	0.529	1.054	0.739	**0.509**
	OSSAT	**0.990**	**0.741**	**0.513**	**1.006**	**0.647**	0.533
	CC(s)	1.174	0.858	0.621	1.299	0.820	0.550
	OSUMC	1.163	0.816	0.637	1.168	0.773	0.638
	UNI	1.592	1.116	0.802	2.215	1.129	0.850

availability of the true outcome labels, the estimates become sufficiently stable, and the difference between methods that incorporate surrogate information become trivial. This observation suggests that OSSAT may be the preferred method when the surrogate outcome is not very accurate and a smaller validation set is preferred.

5 | Discussion

Misclassification of EHR-derived phenotypes can lead to biased and underpowered inference in association studies. Methods using validation samples have been useful in correcting for this bias, but the quality (specifically the informativeness) of the validation set has been given little consideration in the literature. Informative sampling, which uses the information of the outcome, provides an opportunity to improve the quality of the validation set and hence the efficiency of the estimation. When the EHR-derived phenotype is of high accuracy, a simple surrogate-substituted sampling strategy (OSMAC(s)) may be used. However, the surrogate phenotype may also be derived from a simple proxy variable that is only mildly informative of the true outcome (e.g., ICD-10 code or a single biomarker). In this case, a sampling strategy that better utilizes the surrogate phenotype is desired. In this paper, we propose a novel surrogate-assisted sampling strategy (OSSAT) that optimized the statistical efficiency in the estimated model parameters from the validation set, through the use of informative sampling weights. This sampling strategy utilizes the surrogate, EHR-derived phenotype to guide chart review of outcomes in measurement-constrained settings. This surrogate-assisted weight requires a two-step framework, in which a pilot sample with validated outcomes is needed to estimate the accuracy of the surrogate outcome and construct the weights. The OSSAT is proven to optimize the MSE of the overall estimation, according to the same A-optimality criterion [23, 24].

We demonstrate the advantages of the OSSAT method through simulation studies and application to a KPWA data set on secondary breast cancer events (SBCEs), where different levels of surrogate accuracy are evaluated. Both the simulation and real application show that, compared to the simple surrogate-substituted sampling strategy (OSMAC(s)), OSSAT achieves better efficiency in estimating the coefficients, especially when the accuracy of the surrogate outcome is not satisfactory. Moreover, OSSAT is also shown to be robust to differentially misclassified surrogate outcomes.

This novel OSSAT sampling strategy is not without limitations. First, the same as most informative sampling methods, OSSAT also requires a pilot sample to obtain a preliminary estimate of $\hat{\boldsymbol{\beta}}_{MLE}$ that will be used in the optimal weight calculation. Additionally, OSSAT also relies on the pilot sample to estimate the surrogate sensitivity and specificity. The advantage of OSSAT could diminish if the pilot sample is too small in scenarios such as rare outcomes. Second, it is unknown how misspecification of the working models used to construct the weights may impact performance. Our weight formulations suggest that if model misspecification impacts the quantities $\hat{p}(s_i|\boldsymbol{x}_i)$ and $\hat{p}(y_i|s_i, \boldsymbol{x}_i)$, then the resultant weights could be suboptimal from an efficiency standpoint, though $\tilde{\boldsymbol{\beta}}$ should remain unbiased by virtue of inverse probability weighting. Also, the impact of potential differential misclassification of the surrogate is only evaluated empirically. The robustness of the OSSAT method to model misspecification is worthy of future investigation. Third, we use with-replacement sampling to draw validation sample for all the subsampling weights in this paper. This will result in less unique validation sample for chart review. However, as suggested by Wang et al. [32], with-replacement sampling may sacrifice some estimation efficiency compared to Poisson sampling with no replicates. The investigation of applying Poisson sampling with surrogate-assisted sampling weights is beyond the scope of this paper and requires further investigation.

Furthermore, sampling weights based on leverage have the potential to oversample outliers in the data, which could have undue influence on the final estimates. However, as the weighted pseudo log-likelihood uses the inverse of the weights, such influence should be minimal. The use of inverse weighting for estimation is itself a limitation of algorithmic leveraging, as observations with larger weights end up contributing less to the weighted estimation procedure, thus reducing efficiency. [33] improved upon this by proceeding with unweighted estimation and adjusting for bias using a pilot estimate from step 1. The methods discussed in this paper can be modified similarly to further improve efficiency.

Finally, the experimental designs considered in this paper assume the model of interest is chosen as a priori and is representative of the true data generating mechanism. Approaches are needed for the framework of model building and model selection. The literature on active learning may be a useful resource for such extensions.

Overall, we believe this is a useful addition to existing sampling methods for logistic regression models. They may be used alone or in combination with surrogate augmented estimation approaches, such as that proposed by Tong et al. [17], to further improve efficiency. An R package for implementing the proposed weights is forthcoming.

Acknowledgments

This work was supported in part by National Institutes of Health (R21AI167418, 1R01LM014344, 1R01AG077820, R01LM012607, R01AI1-30460, R01AG073435, R56AG074604, R01LM013519, R56AG069880, U01TR003709, RF1G077820, R21CA227613, R21CA143242, R01CA-09377, U01CA063731, and R01CA120562). This work was supported partially through Patient-Centered Outcomes Research Institute (PCORI) Project Program Awards (ME-2019C3-18315 and ME-2018C3-14899). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee, or the National Institutes of Health.

Disclosure

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from Kaiser Permanente Washington. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of Kaiser Permanente Washington.

References

1. G. Hripcsak and D. J. Albers, "Next-Generation Phenotyping of Electronic Health Records," *Journal of the American Medical Informatics Association* 20, no. 1 (2012): 117–121.

2. P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care," *Nature Reviews Genetics* 13, no. 6 (2012): 395–405.

3. C. B. Forrest, P. A. Margolis, L. C. Bailey, et al., "PEDSnet: A National Pediatric Learning Health System," *Journal of the American Medical Informatics Association* 21, no. 4 (2014): 602–606.

4. Y. Wu, J. L. Warner, L. Wang, et al., "Discovery of Noncancer Drug Effects on Survival in Electronic Health Records of Patients With Cancer: A New Paradigm for Drug Repurposing," *JCO Clinical Cancer Informatics* 3 (2019): 1–9.

5. M. E. Menendez, S. J. Janssen, and D. Ring, "Electronic Health Record-Based Triggers to Detect Adverse Events After Outpatient Orthopaedic Surgery," *BMJ Quality and Safety* 25, no. 1 (2016): 25–30.

6. J. C. Kirby, P. Speltz, L. V. Rasmussen, et al., "PheKB: A Catalog and Workflow for Creating Electronic Phenotype Algorithms for Transportability," *Journal of the American Medical Informatics Association* 23, no. 6 (2016): 1046–1052.

7. R. A. Hubbard, J. Huang, J. Harton, et al., "A Bayesian Latent Class Approach for EHR-Based Phenotyping," *Statistics in Medicine* 38, no. 1 (2019): 74–87.

8. L. Wang, J. E. Olson, S. J. Bielinski, et al., "Impact of Diverse Data Sources on Computational Phenotyping," *Frontiers in Genetics* 11 (2020): 556.

9. S. Denaxas, K. Direk, A. Gonzalez-Izquierdo, et al., "Methods for Enhancing the Reproducibility of Biomedical Research Findings Using Electronic Health Records," *Biodata Mining* 10, no. 1 (2017): 31.

10. R. Duan, M. Cao, Y. Wu, et al., "An Empirical Study for Impacts of Measurement Errors on EHR Based Association Studies," in *Proceedings of the American Medical Informatics Association Annual Symposium* (American Medical Informatics Association, 2016), 1764.

11. Y. Chen, J. Wang, J. Chubak, and R. A. Hubbard, "Inflation of Type I Error Rates due to Differential Misclassification in EHR-Derived Outcomes: Empirical Illustration Using Breast Cancer Recurrence," *Pharmacoepidemiology and Drug Safety* 28, no. 2 (2019): 264–268.

12. J. M. Neuhaus, "Bias and Efficiency Loss due to Misclassified Responses in Binary Regression," *Biometrika* 86, no. 4 (1999): 843–855.

13. R. A. Hubbard, J. Tong, R. Duan, and Y. Chen, "Reducing Bias due to Outcome Misclassification for Epidemiologic Studies Using EHR-Derived Probabilistic Phenotypes," *Epidemiology* 31, no. 4 (2020): 542–550.

14. S. Martin, J. Wagner, N. Lupulescu-Mann, et al., "Comparison of EHR-Based Diagnosis Documentation Locations to a Gold Standard for Risk Stratification in Patients With Multiple Chronic Conditions," *Applied Clinical Informatics* 8, no. 3 (2017): 794–809.

15. A. Chakrabortty and T. Cai, others, "Efficient and Adaptive Linear Regression in Semi-Supervised Settings," *Annals of Statistics* 46, no. 4 (2018): 1541–1572, https://doi.org/10.1214/17-AOS1594.

16. D. Cheng, A. N. Ananthakrishnan, and T. Cai, "Robust and Efficient Semi-Supervised Estimation of Average Treatment Effects With Application to Electronic Health Records Data," *Biometrics* 77, no. 2 (2020): 413–423.

17. J. Tong, J. Huang, J. Chubak, et al., "An Augmented Estimation Procedure for EHR-Based Association Studies Accounting for Differential Misclassification," *Journal of the American Medical Informatics Association* 27, no. 2 (2020): 244–253.

18. P. Ma, M. W. Mahoney, and B. Yu, "A Statistical Perspective on Algorithmic Leveraging," *Journal of Machine Learning Research* 16, no. 1 (2015): 861–911.

19. P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Sampling Algorithms for $\ell 2$ Regression and Applications," in *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms* (Society for Industrial and Applied Mathematics, 2006), 1127–1136.

20. M. W. Mahoney and P. Drineas, "CUR Matrix Decompositions for Improved Data Analysis," *Proceedings of the National Academy of Sciences* 106, no. 3 (2009): 697–702.

21. P. Ma and X. Sun, "Leveraging for Big Data Regression," *Wiley Interdisciplinary Reviews: Computational Statistics* 7, no. 1 (2015): 70–76.

22. W. Fithian and T. Hastie, "Local Case-Control Sampling: Efficient Subsampling in Imbalanced Data Sets," *Annals of Statistics* 42, no. 5 (2014): 1693–1724.

23. H. Wang, R. Zhu, and P. Ma, "Optimal Subsampling for Large Sample Logistic Regression," *Journal of the American Statistical Association* 113, no. 522 (2018): 829–844.

24. T. Zhang, Y. Ning, and D. Ruppert, "Optimal Sampling for Generalized Linear Models Under Measurement Constraints," *Journal of Computational and Graphical Statistics* 30, no. 1 (2021): 106–114.

25. M. S. Pepe, M. Reilly, and T. R. Fleming, "Auxiliary Outcome Data and the Mean Score Method," *Journal of Statistical Planning and Inference* 42, no. 1–2 (1994): 137–160.

26. N. Chan, "A-Optimality for Regression Designs," *Journal of Mathematical Analysis and Applications* 87, no. 1 (1982): 45–50.

27. D. Firth, "Bias Reduction of Maximum Likelihood Estimates," *Biometrika* 80, no. 1 (1993): 27–38.

28. D. M. Boudreau, O. Yu, J. Chubak, et al., "Comparative Safety of Cardiovascular Medication Use and Breast Cancer Outcomes Among Women With Early Stage Breast Cancer," *Breast Cancer Research and Treatment* 144, no. 2 (2014): 405–416.

29. J. Chubak, O. Yu, G. Pocobelli, et al., "Administrative Data Algorithms to Identify Second Breast Cancer Events Following Early-Stage Invasive Breast Cancer," *Journal of the National Cancer Institute* 104, no. 12 (2012): 931–940.

30. L. Mellemkjær, S. Friis, J. H. Olsen, et al., "Risk of Second Cancer Among Women With Breast Cancer," *International Journal of Cancer* 118, no. 9 (2006): 2285–2292.

31. Y. Cheng, Z. Huang, Q. Liao, et al., "Risk of Second Primary Breast Cancer Among Cancer Survivors: Implications for Prevention and Screening Practice," *PLoS One* 15, no. 6 (2020): e0232800.

32. J. Wang, J. Zou, and H. Wang, "Sampling With Replacement vs Poisson Sampling: A Comparative Study in Optimal Subsampling," *IEEE Transactions on Information Theory* 68, no. 10 (2022): 6605–6630.

33. H. Wang, "More Efficient Estimation for Logistic Regression With Optimal Subsamples," *Journal of Machine Learning Research* 20 (2019): 1–59.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1.** Web appendix. **Data S2.** Supporting information 1. **Data S3.** Supporting information 2.